Q2

Hun Kang

2024-09-16

Q2.

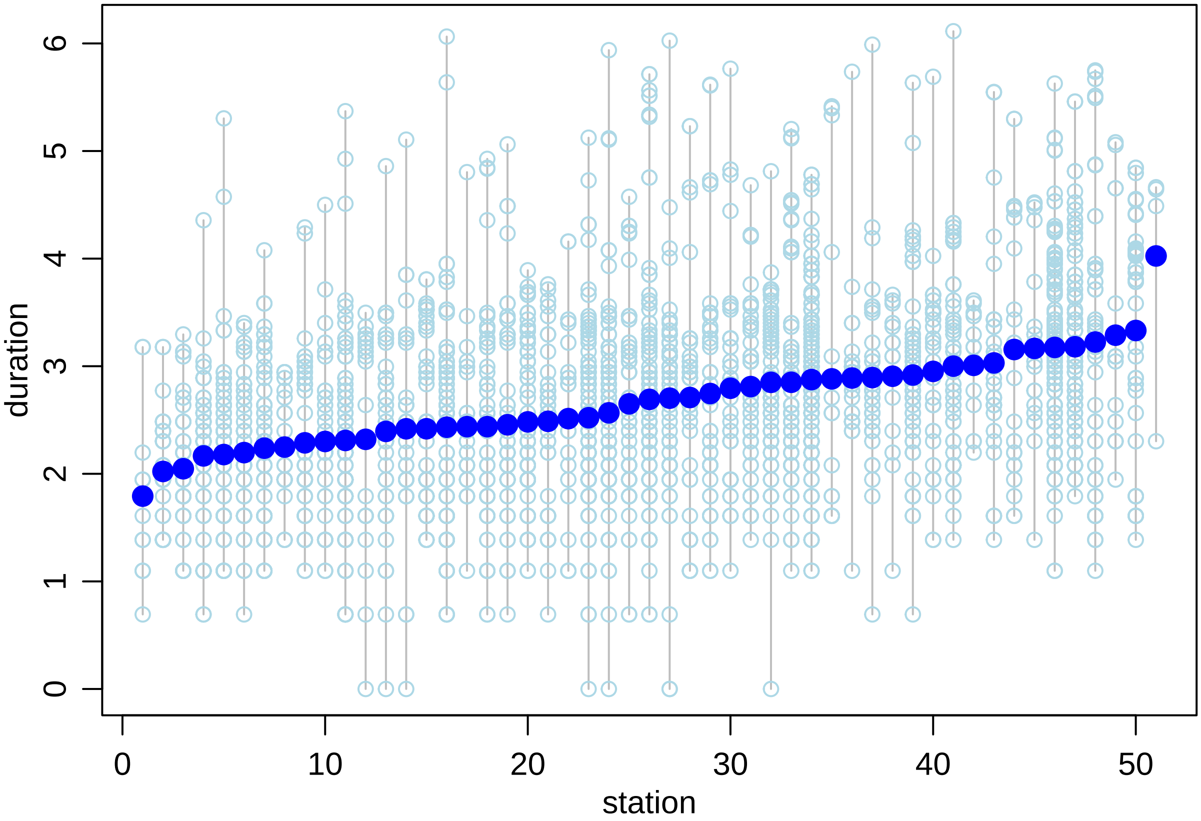
```
dat = readRDS(url("https://www2.stat.duke.edu/~pdh10/Teaching/610/Homework/btrips2015-7-1-4.rds"))
dat$station = as.factor(dat$station)
str(dat)
```

```
## ’data.frame’:    2526 obs. of  2 variables:
##  $ station : Factor w/ 51 levels "BT-01","BT-03",..: 31 34 30 39 18 39 17 33 33 12 ...
##  $ duration: num  4 3 16 14 5 12 9 5 4 14 ...
```

```
gdotplot<-function(y,g,xlab="group",ylab="response",mcol="blue",
                   ocol="lightblue",sortgroups=TRUE,...)
{
  m<-length(unique(g))
  rg<-rank( tapply(y,g,mean),ties.method="first")
  if(sortgroups==FALSE){ rg<-1:m ; names(rg)<-unique(g)}
  plot(c(1,m),range(y),type="n",xlab=xlab,ylab=ylab)

  for(j in unique(g))
  {
    yj<-y[g==j]
    rj<-rg[ match(as.character(j),names(rg)) ]
    nj<-length(yj)
    segments(rj ,max(yj),rj,min(yj),col="gray")
    points( rep(rj,nj), yj,col=ocol, ...)
    points(rj,mean(yj),pch=16,cex=1.5,col=mcol)
  }
}

par(mar=c(3,3,1,1), mgp=c(1.75,.75,0))
gdotplot(log(dat$duration),
         dat$station,
         xlab="station", ylab="duration")
```



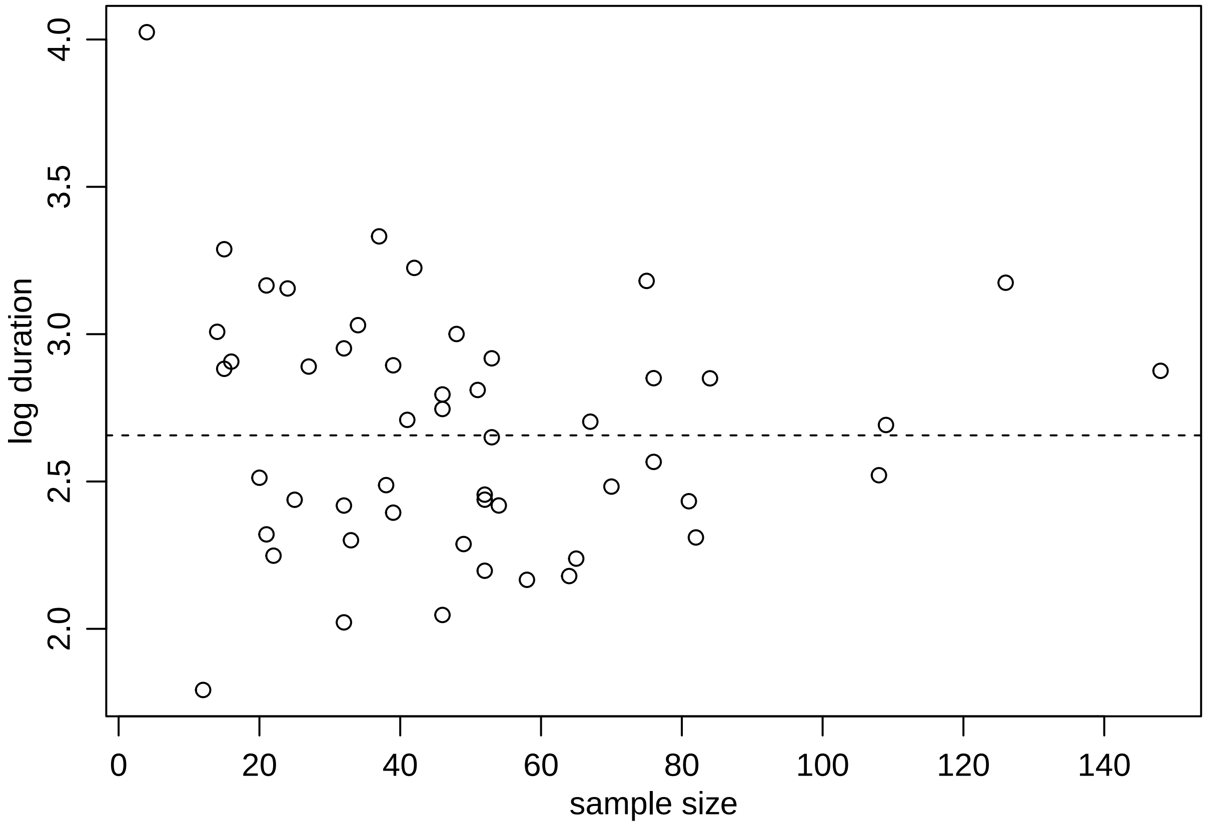
a.

```
anova(lm(log(duration) ~ station, data = dat))
```

```
## Analysis of Variance Table
##
## Response: log(duration)
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## station     50  289.86  5.7971  7.6989 < 2.2e-16 ***
## Residuals 2475 1863.62  0.7530
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b.

```
njs = aggregate(log(duration) ~ station, data = dat, length)[,2]
yjbars = aggregate(log(duration) ~ station, data = dat, mean)[,2]
par(mar=c(3,3,1,1), mgp=c(1.75,.75,0))
plot(njs, yjbars, xlab= "sample size", ylab= "log duration")
abline(h = mean(log(dat$duration)), lty=2)
```



c.

```r
library(lme4)
```

```r
mod <- lmer(log(duration) ~ 1 + (1 | station), data = dat, REML = FALSE)
summary(mod)
```

```
## Linear mixed model fit by maximum likelihood  [’lmerMod’]
## Formula: log(duration) ~ 1 + (1 | station)
##    Data: dat
## 
##      AIC      BIC   logLik deviance df.resid 
##   6563.3   6580.8  -3278.6   6557.3     2523 
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -3.2654 -0.6489 -0.0511  0.5580  4.1608 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  station  (Intercept) 0.1119   0.3345  
##  Residual             0.7542   0.8685  
## Number of obs: 2526, groups:  station, 51
```

```
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   2.6576     0.0511      52
```

We can extract the estimates of μ, τ and σ as below. Here, the estimate of σ is a pooled estimate.

```
fixef(mod)
```

```
## (Intercept)
##    2.657586
```

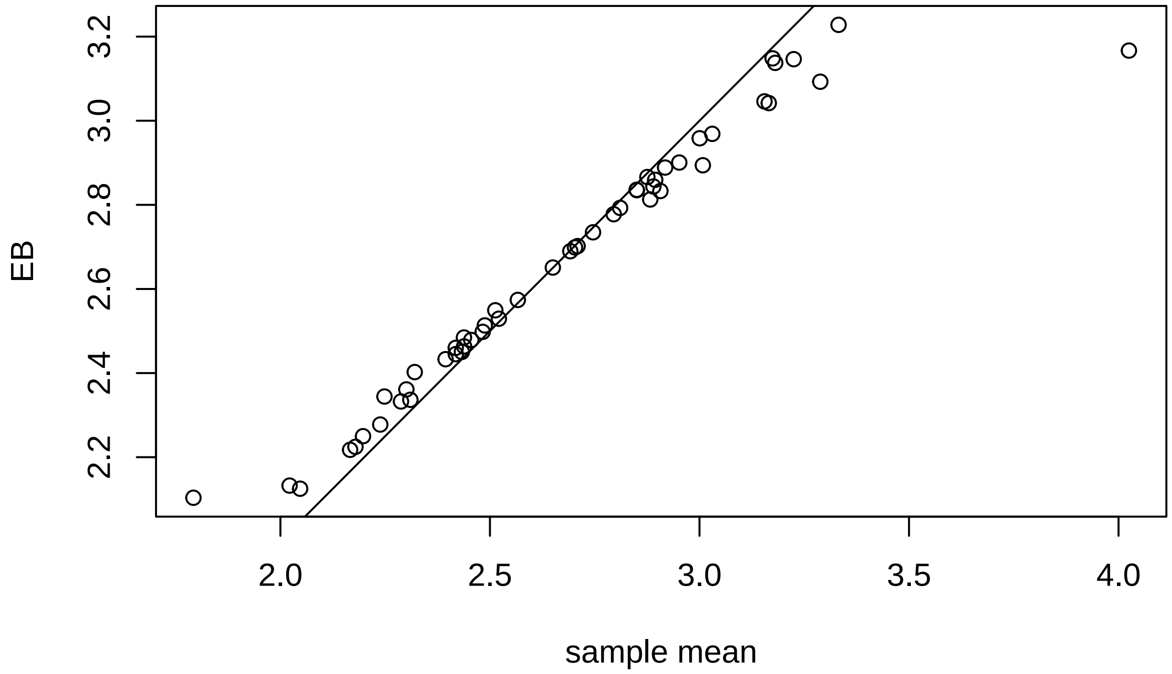
```
data.frame(VarCorr(mod))
```

```
##        grp        var1 var2      vcov     sdcor
## 1  station (Intercept) <NA> 0.1119095 0.3345288
## 2 Residual        <NA> <NA> 0.7542276 0.8684628
```

d.

$$\hat{\theta}_j = \frac{n_j/\hat{\sigma}^2}{n_j/\hat{\sigma}^2 + 1/\hat{\tau}^2}\bar{y}_j + \frac{1/\hat{\tau}^2}{n_j/\hat{\sigma}^2 + 1/\tau^2}\hat{\mu}$$

```
muhat = fixef(mod)
sig2hat = data.frame(VarCorr(mod))$vcov[2]
tau2hat = data.frame(VarCorr(mod))$vcov[1]
thetahat = function(yjbars, njs, muhat, sig2hat,tau2hat){
  w = (njs/sig2hat) / (njs/sig2hat + 1/tau2hat)
  w * yjbars + (1-w) * muhat
}
thetahats = thetahat(yjbars, njs, muhat, sig2hat, tau2hat)
plot(yjbars, thetahats, xlab = "sample mean", ylab = "EB")
abline(a=0,b=1)
```



e.

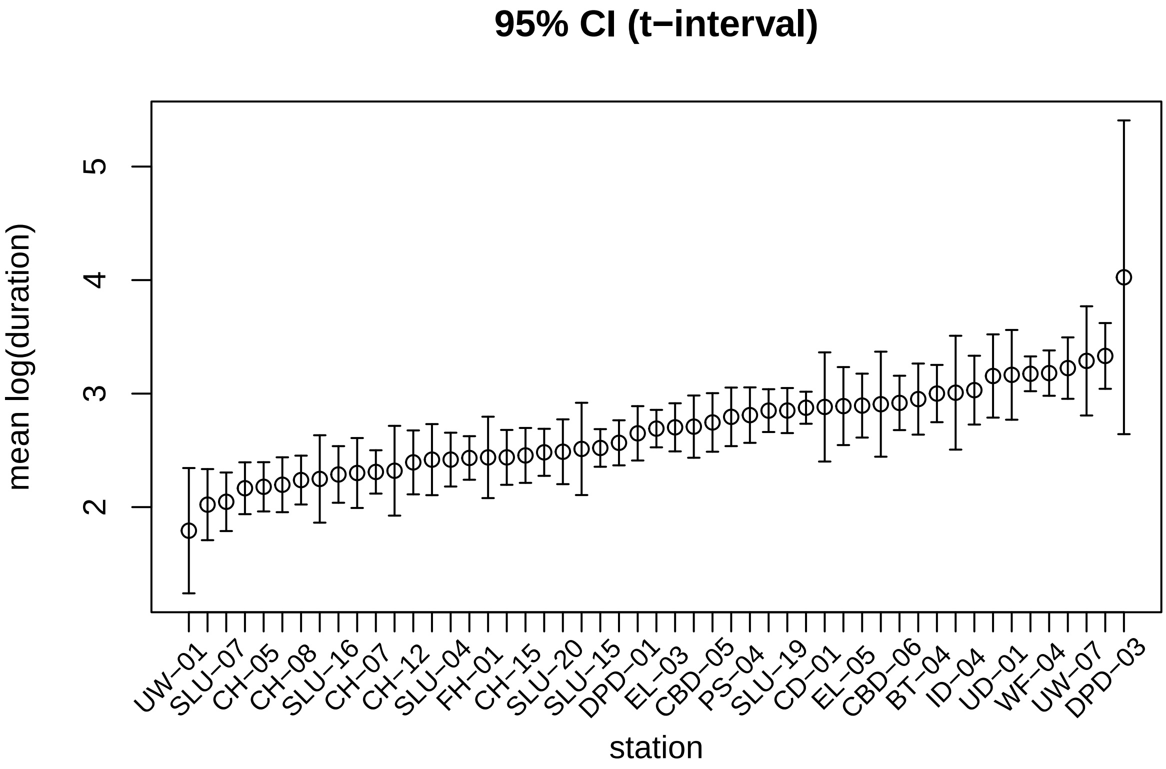
$$\bar{y}_j \pm \frac{t_{n_j-1,1-\alpha/2}}{\sqrt{n_j/\hat{\sigma}^2}}$$

```
library(tidyverse)
```

```
yjbars_ci = aggregate(log(duration) ~ station, data = dat, mean)
yjbars_ci$nj = njs
yjbars_ci = yjbars_ci %>%
  mutate(lb = `log(duration)` - qt(1-0.025, nj-1)* sqrt(sig2hat/nj),
         ub = `log(duration)` + qt(1-0.025, nj-1)* sqrt(sig2hat/nj)) %>%
  arrange(`log(duration)`)
head(yjbars_ci)
```

```
##   station log(duration) nj       lb       ub
## 1   UW-01      1.792443 12 1.240648 2.344237
## 2   CH-09      2.021864 32 1.708749 2.334978
## 3  SLU-07      2.047065 46 1.789163 2.304966
## 4   CH-01      2.166367 58 1.938016 2.394717
## 5   CH-05      2.179053 64 1.962118 2.395989
## 6  SLU-01      2.197129 52 1.955347 2.438911
```

```r
plot(1:nrow(yjbars_ci), yjbars_ci$`log(duration)`,
     ylim = range(yjbars_ci[, c("lb", "ub")]),
     xaxt = "n",
     xlab = "station", ylab = "mean log(duration)",
     main = "95% CI (t-interval)")
idx = seq(1, nrow(yjbars_ci), by=2)
axis(1, at = (1:nrow(yjbars_ci)), labels=F)
text(cex=0.8, x=(1:nrow(yjbars_ci))[idx]-1, y=0.5, yjbars_ci$station[idx], xpd=TRUE, srt=45)
arrows(1:nrow(yjbars_ci), yjbars_ci$lb,
         1:nrow(yjbars_ci), yjbars_ci$ub,
       code=3, angle=90, length=0.03)
```


95% CI (t−interval)
mean log(duration)
UW−01
SLU−07
CH−05
CH−08
SLU−16
CH−07
CH−12
SLU−04
FH−01
CH−15
SLU−20
SLU−15
DPD−01
EL−03
CBD−05
PS−04
SLU−19
CD−01
EL−05
CBD−06
BT−04
ID−04
UD−01
WF−04
UW−07
DPD−03
station