hw5_q.R

March 3, 2019

```
library(hexbin) # load library
### 2
round(Q_dist, 3) # dist of digits for entire dataset

### 3
BenFord <- function(x) { # Benford's theoretical distribution
  log10(1+1/x)
}

# Plot Benford with actual distribution
plot(BenFord(1:9), ylab = 'Density', xlab = 'Digit',
     main = 'Benford\'s Theoretical Distribuion with Actual Distribution')
legend(x = 5, y = 0.25, legend = c('Benford', 'Actual'), col = c('blue', 'red'), lty = 1:3)
par(pch=20, col="blue"); lines(BenFord(1:9)); points(BenFord(1:9))
par(pch=22, col="red"); points(prop.table(comb_sum)); lines(prop.table(comb_sum), type = 'c')
dev.off()

# Calculate KLD for actual vs Benford, uniform vs Benford
KLD(P = Q_dist, Q = BenFord(1:9))
KLD(P = rep(1/9, 9), Q = BenFord(1:9))

### 4
length(grouped_id) # Number of different id's
View(rownames(comb_dist)) # Inspect strange id's

# Summary statistics (mean, median, mode, etc.)
summary(KLD_df)
unique(KLD_df)[which.max(tabulate(match(KLD_df, unique(KLD_df))))] # mode
alpha <- 0.05; quantile(KLD_df, c(alpha/2, 1-alpha/2))
hist(KLD_df, main = 'Histogram of KLD scores', xlab = 'KLD') # histogram

### 5
KLD_df[KLD_df > 2.5,] # how many larger than 2.5

### bootstrap
KLD_ci_final <- read.csv('KLD_ci.csv') # Load data from cluster
colnames(KLD_ci_final) <- c('id', 'actual', '2.5', '97.5')

ci_width <- KLD_ci_final[,c('97.5')] - KLD_ci_final[,c('2.5')] # Calculate CI width
summary(ci_width)
quantile(ci_width, c(alpha/2, 1-alpha/2))
unique(ci_width)[which.max(tabulate(match(ci_width, unique(ci_width))))] # mode
hist(ci_width, main = 'Histogram of C.I. Width', xlab = 'C.I. Width')

# Check upper and lower bounds of bootstrap
upper_bound <- KLD_ci_final$`97.5` - KLD_ci_final$actual
upper_bound[upper_bound < 0] # None beyond upper bound
```

```r
lower_bound <- KLD_ci_final$actual - KLD_ci_final$`2.5`
length(lower_bound[lower_bound < 0]) / length(KLD_df) # 104 recipients below lower bound

# width as a function of size
width_size <- data.frame(ci_width, id_sizes)
width_size$id <- rownames(width_size)
rownames(width_size) <- NULL
colnames(width_size) <- c('bootstrap_widths', 'id_lengths', 'ids')

nonzero_ci <- width_size[width_size$bootstrap_widths > 0,]
nonzero_ci <- nonzero_ci[-1,] # ignore NA recipient id
nonzero_ci[,1:2] <- log(nonzero_ci[,1:2]) # log transform

hex_plot <- hexbin(nonzero_ci$id_lengths, nonzero_ci$bootstrap_widths) # Plot
P = plot(hex_plot, main = 'Hexbin log C.I. Widths Regressed on log Recipient Lengths',
         xlab = 'log Recipient Lengths', ylab = 'log C.I. Widths')
fit <- lm(bootstrap_widths ~ id_lengths, data = nonzero_ci)
hexVP.abline(hvp = P$plot.vp, a = fit$coefficients[1], b = fit$coefficients[2])
```