STAT 391 HW7

Chongyi Xu

May 28, 2018

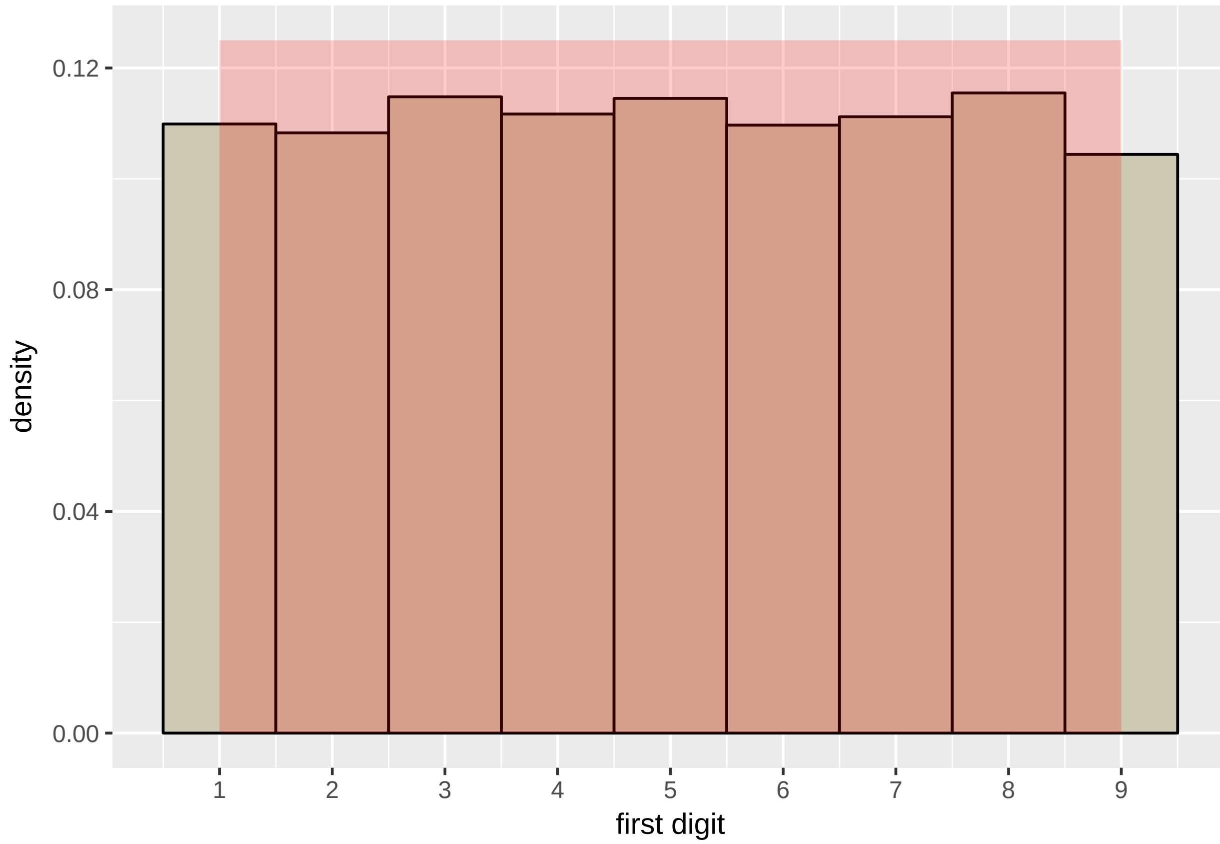
Problem 1 - Testing a hypothesis

a. If positive integer numbers with at most 3 digits are drawn uniformly, show that the distribution of the first digit is uniform over $S = \{1, \cdots, 9\}$.

```
set.seed(391)
n <- 10000
x <- runif(n, min=1, max=999)

fdigit <- function(x) {
as.numeric(head(
  strsplit(as.character(x),'')[[1]],n=1
  ))
}

d <- data.frame(first=sapply(x,fdigit))
xx <- seq(from=1, to=9, length=10000)
u <- dunif(xx, min=1, max=9)

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

```
ggplot(d, aes(first)) + geom_histogram(aes(y=..density..),
                        col='black', fill='cornsilk3',bins=9) +
geom_area(aes(x=xx, y=u), alpha=0.2, fill='red') +
scale_x_continuous(name='first digit', breaks=c(1,2,3,4,5,6,7,8,9))
```



From the plot, we can see that the distribution of first digits from simulated data follows the uniform distribution pretty well.

See detailed proof in part (b).

b. Prove if positive integer numbers with at most d digits are drawn uniformly, then the distribution of the first digit is uniform over $S = \{1, \cdots, 9\}$

Prove using induction. Denote the positive number as x

- Base case ($d = 1$)

$x \sim U(1, 9)$, the first digit is trivially uniform over S.

- Induction (assume when $d = k$, the distribution of the first digit is uniform over S, show $d = k + 1$ also works)

From assumption we have $x \sim U(1, \sum_{i=0}^{k-1} 10^i)$, then $10x + x \sim U(10, \sum_{i=1}^{k} 10^i) + U(1, 9) \sim U(1, \sum_{i=1}^{k} 10^i)$, which is the distribution of x at $d = k + 1$. Q.E.D.

c. Consider the event $E_{n,t}$ =“in a data set of n integers, at least t of them start with 1”. Write an expression $p_{n,t} = P_0(E_{n,t})$, the probability that E_n is true given that the highest digits are uniformly distributed over S. This should be a function of t and n.

Since at part(b), we have proved that if positive numbers with at most d digits are drawn uniformly, then the distribution of the first digit is uniform over $S = \{1, \cdots, 9\}$. Then $p_{n,t}$ can be interpreted as $P\{\sum_{i=1}^{n} I(x_i = 1) \geq t | x_i \sim U(1, 9)\}$. We can see this is just the probability of a binomial random variable with $p = \frac{1}{9}$. Denote Y Binom$(n, p = \frac{1}{9})$Therefore we have

$$Pr\{\sum_{i=1}^{n} I(x_i = 1) \geq t | x_i \sim U(1,9)\} = Pr\{Y \geq t\}$$

$$= 1 - \sum_{k=1}^{t-1} \binom{n}{k} (\frac{1}{9})^k (\frac{8}{9})^{n-k}$$

d. Read the first $n_D = 60$ data from file `hw6_digit.dat`. Compute p_{n_D,t_D} from the data.

```
dat <- readLines('hw6_digits.dat')[1:60]
fd <- sapply(dat,fdigit)
nd <- length(dat)
td <- length(which(fd==1)) - 1
1 - sum(choose(nd, 1:td) * (1/9)^(1:td) * (8/9)^(nd-1:td))
```

```
## [1] 0.03154641
```

```
td <- 6-1
1 - sum(choose(nd, 1:td) * (1/9)^(1:td) * (8/9)^(nd-1:td))
```

```
## [1] 0.6693548
```

g. We again only use the first 60 data. Now we consider another way of testing. Denote $\theta_i = P(\text{first digit is } i)$. Let model A be that the first digit follows a uniformly distribution of over S. Let B be that the first digit tfollows a multinomial distribution over S.

Compute the likelihood of the data under model A. Compute the ML estimates $\hat{\theta}_i^B$ for $i = 1, \cdots, 9$ under model B, and then use them to obtain the maximum likelihood of the data under model B. Use these two quantitites to obtain the likelihood ratio test statistics value λ_D.

```
dat <- readLines('hw6_digits.dat')[1:60]
fd <- sapply(dat,fdigit)
n <- length(dat)
ni <- rep(NA, 9)
thetaB <- rep(NA, 9)
for (i in 1:9) {
ni[i] <- length(which(fd==i))
thetaB[i] <- ni[i]/n
}
L_A <- (1/9)^n
L_B <- exp(lfactorial(n) - sum(lfactorial(ni)) +
           sum(ni * log(thetaB)))
lambdaD <- L_A/L_B
print(paste('lambda_D=', lambdaD))
```

```
## [1] "lambda_D= 2.32058115821921e-52"
```

h. How many free parameters d_B are estimated from data in model B? Use the χ^2 table to obtain $Pr[Z_d > -2ln\lambda_D]$ where Z_d is a random variable drawn from a χ^2 distribution with $d = d_B - d_A$ degrees of freedom.

```
dB <- 9-1
dA <- 0
d <- dB - dA
pr <- pchisq(-2*log(lambdaD), df=d, lower.tail=F)
print(paste('pr=', pr))
```

```
## [1] "pr= 6.66674523706425e-47"
```

i. Now read the whole data in `hw6-digit.dat` and compute the above probability for the whole data set.

```
thetaB <- rep(NA, 9)
dat <- readLines('hw6_digits.dat')
fd <- sapply(dat,fdigit)
n <- length(dat)
ni <- rep(NA, 9)
for (i in 1:9) {
ni[i] <- length(which(fd==i))
thetaB[i] <- ni[i]/n
}
L_A <- exp(n*log(1/9))
L_B <- exp(lfactorial(n) - sum(lfactorial(ni)) +
           sum(ni * log(thetaB)))
lambdaD <- L_A/L_B
print(paste('lambda_D=', lambdaD))
```

```
## [1] "lambda_D= 0"
```

```
pr <- pchisq(-2*log(lambdaD), df=d, lower.tail=F)
print(paste('pr=', pr))
```

```
## [1] "pr= 0"
```

With a significant level of 0.001, we reject our null hypothesis that $\theta = \theta^A$, that is, the data set is not drawn from uniform distribution.

Problem 2 - Rob at the Flintstone factory

a. Let Y denote the length measurement of flintstone. Under the model that all flintstones are exactly l_0 long, what is the distribution of Y? What are the $E[Y]$ and $Var(Y)$?

Since we have known that error term $X \sim U(-1, 1)$, and $Y = 8 + X$, then $Y \sim U(7, 9)$.

Thus $E[Y] = 8$, $Var(Y) = \frac{1}{12}(9-7)^2 = \frac{1}{3}$

b. Under the model above, what is the sample space of this variable? What is $E[L]$? What is $Var(L)$?

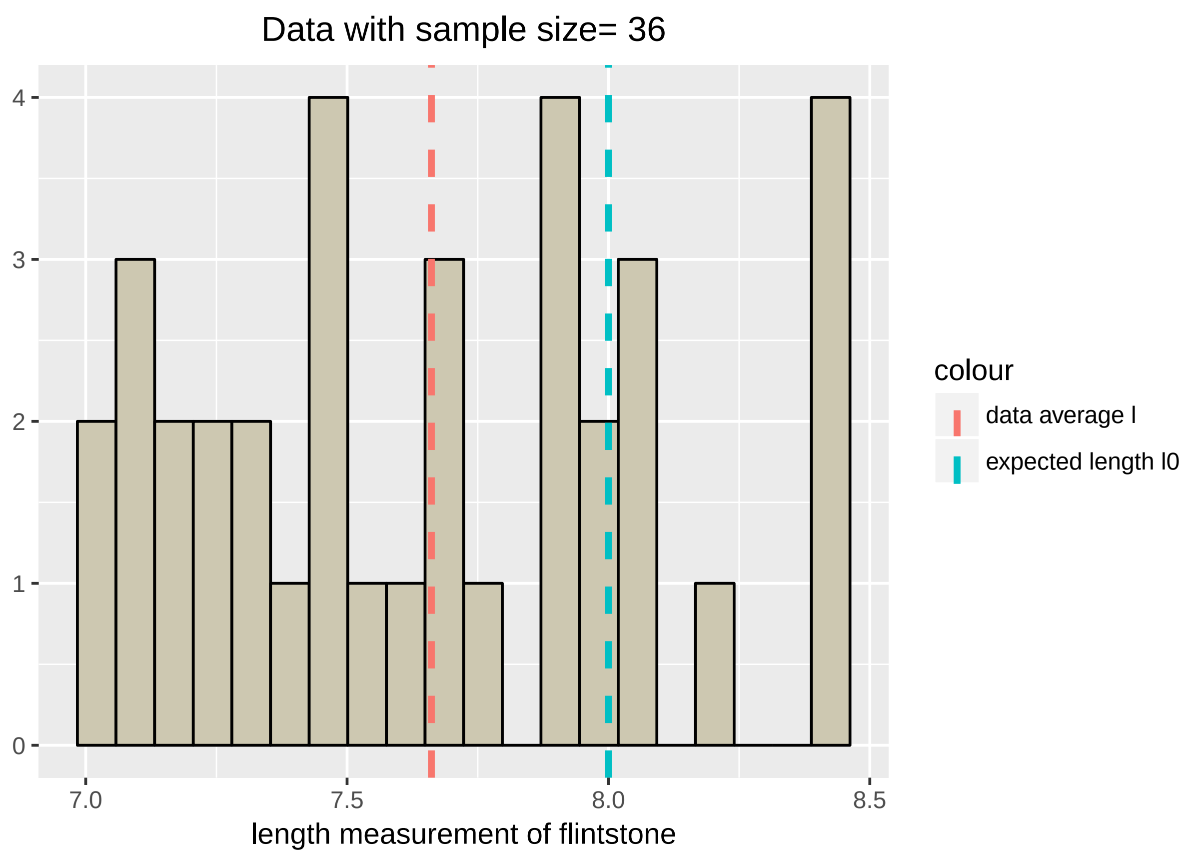
Since we have $Y \sim U(7, 9)$ in part(a), L then has the distribution that $L = \frac{1}{n}\sum_{i=1}^{n} Y_i \sim \frac{1}{n}\sum U(7, 9)$. So the sample space of L is $[7, 9]$ and $E[L] = 8$, $Var(L) = \frac{1}{3n}$

c. The acutal measurements made by Rob are in `flintstones.dat`. Make a plot of the data, also marking clearly the sample size S_Y, the point l_0 and the point $L = l$ the data average.

```
flintstones <- scan('flintstones_1.dat', sep=' ')
l <- mean(flintstones)
l0 <- 8
Sy <- length(flintstones)
```

```
ggplot() + geom_histogram(aes(flintstones),
                         col='black',
                         fill='cornsilk3', bins=20) +
geom_vline(aes(xintercept=l, col='data average l'),
           lwd=1.2, linetype='dashed') +
geom_vline(aes(xintercept=l0, col='expected length l0'),
           lwd=1.2, linetype='dashed') +
xlab('length measurement of flintstone') +
```

```r
ylab('') + ggtitle(paste('Data with sample size=', Sy)) +
theme(plot.title = element_text(hjust = 0.5))
```



d. Rob decides to use Chebyshev's inequality.

$$Prob[|z - E[Z] \geq t] \leq \frac{Var(Z)}{t^2}$$

Apply this inequality to the variable L; assuming that Fred says the truth, L should have the mean and variance you obtained in b. Therefore, the inequality will tell how probable it is for the actual $L = l$ Rob have calculated from the data to occur. Denote this probility p_{Cheb}.

$$\begin{aligned} Prob[|z - E[Z]| \geq t] &= Prob[|l - 8| \geq t] \\ &\leq \frac{Var(Z)}{t^2} \\ &= \frac{1}{3t^2} \end{aligned}$$

Therefore, $Prob[l - 8 \geq t] \leq \frac{1}{3t^2}$ for all $t > 0$. In this question, we are considering $|l - l_0| \leq 1$ from the problem 2 statement. So

```r
t <- 1
length(which(abs(flintstones-mean(flintstones))>=t))/Sy
```

```
## [1] 0
```

```
pCheb <- 1/(3*Sy*t^2)
pCheb
```

```
## [1] 0.009259259
```

e. Rob now wants to use a more refined tool. He knows about the CLT.

```
zn <- (sum(flintstones)-Sy*8)/sqrt(Sy*1/3)
pr <- pnorm(zn, mean=0, sd=1)
print(paste('pr=', pr))
```

```
## [1] "pr= 0.000215149017390636"
```

g. In addition to the probability $p = p_<$, Rob also computed the probability $p_>$, and the probability $p_{\neq}$.

Write these quantites as probability statements involving Z and z_n and find the numerical values of $p_<$ and $p_{\neq}$

$$
\begin{aligned}
p_< &= Pr[Z < z_n] \\
p_> &= Pr[Z > z_n] \\
p_{\neq} &= Pr[Z > |z_n|] + Pr[Z < -|z_n|]
\end{aligned}
$$

```
prNoLarger <- pr
prNoSmaller <- pnorm(zn, mean=0, sd=1, lower.tail=F)
prAbs <- pnorm(abs(zn), lower.tail=F) + pnorm(-abs(zn))
print(paste('p_> = ', prNoSmaller))
```

```
## [1] "p_> =  0.999784850982609"
```

```
print(paste('p_neq = ', prAbs))
```

```
## [1] "p_neq =  0.000430298034781272"
```

h. Let H_0 be Fred's claim that the flintstones are sampled from the true model in a. Let H_1 be the alternative that the flintstones are sampled from a uniform distribution with lower mean. Compute the Max Likelihood Estimator for the data under the alternative model.

```
a0 <- 7
b0 <- 9
a1 <- min(flintstones)
b1 <- max(flintstones)
theta_MLE <- b1-a1
print(paste('theta_MLE = ', theta_MLE))
```

```
## [1] "theta_MLE =  1.4042"
```

i. Now compute the likelihood ratio λ and the test statistics $t = -2ln\lambda$. How many free parameters has the model H_1? Let this number be d.

Since H_1 is also a uniform distribution model, $d_1 = 2$ due to independence of a_1 and b_1.

```
lambda <- ((1/(b0-a0))/(1/theta_MLE))^Sy
print(paste('lambda=', lambda))
```

```
## [1] "lambda= 2.95367719073439e-06"
```

```
pr <- pchisq(-2*log(lambda), df=2, lower.tail=F)
print(paste('p-value=', pr))
```

```
## [1] "p-value= 2.95367719073439e-06"
```

j. How do you explain the difference between p_{Cheb} and p_{LR}.

Comapring the value of p_{Cheb} with p_{LR}, I found that at significant level $\alpha = 0.001, 0.005$, only the likelihood ratio test rejects the null hypothesis that the flintstones are sampled from the true model in a. At significant level $\alpha = 0.01, 0.05, 0.1 \cdots$, both tests reject the null hypothesis.