Computing assignments (Spring 2016)

Homework 3

Due: 8 pm, Feb. 2

Each of you receive 500 numbers, denoted as $X_1,\dots,X_{500}$, all of which follow a normal distribution with an **unknown** mean and an **unknown** variance. **Please read following questions carefully. Note that not all numbers will be used!!!**

The goals include finding a point estimator and a confidence interval for μ with good accuracy.

1. (1pt) Choose the first 10 observations, $X_1,\dots,X_{10}$, as your sample. Treat this as a pilot sample (an experimental and preliminary sample.) Report an estimate of the unknown population variance. In R, the following command extracts the first 10 elements of vector `x` and save them as a new vector `y`.`y=x[1:10]` Recall that `sum( (y-mean(y))^2 )/(length(y)-1)` gives your the sample variance. So does `var(y)`. Check both commands to see if they match.
2. (2pts) Pretend that 10 is a large number (although it is not that large, let's pretend it that way for now.) Construct a 90% confidence interval using only the first 10 observations (the sample mean is based on the 10 observations, and the sample standard deviation is also based on the 10 observations.) Report the two-sided confidence interval: lower bound in (2a) and upper bound in (2b). For this question, ignore the materials in Section 8.8. That is, please use the formula in Section 8.6. To find the percentile point of a standard normal distribution, do not use the SOA normal table or Table 4 in the text book. Instead, you can use command `z=qnorm(0.995)`, `z=qnorm(0.975)`, `z=qnorm(0.95)`, etc, to find the percentile point. This request is to help us grade your answer numerically. For example, `qnorm(0.975)` gives 1.959964, which corresponds to 1.96 in the normal tables. If you use 1.96 instead of 1.959964, your answer may be mistakenly graded as incorrect. Type in `?qnorm` in R for more information on the function `qnorm()`.
3. (2pts) Redo the last question except that you use the formula in Section 8.8, that is, provide a 90% confidence interval for a small sample. To find the percentile point of a t distribution, do not use Table 5 in the text book. Instead, you can use command `t=qt(0.995,9)`, `z=qt(0.975,9)`, `z=qt(0.95,9)`, etc, to find the percentile point. The first argument is the left-tail (not right-tail) probability and the second argument is the degrees of freedom (which is 9 here). For example, `qt(0.995,9)` gives 3.249836, which corresponds to 3.250 in Table 5, row 9, last column. Compare `qt(0.995,9)`, `qt(0.99,9)`, `qt(0.975,9)`, or `qt(0.95,9)` with row 9 of Table 5. Type in `?qt` in R for more information on the function `qt()`.
4. (2pts) The confidence interval obtained above does not provide a lot of information about μ since it is too wide. In addition, the sample mean based on only 10 observations is also unlikely to be accurate. **In this problem, we want to find an estimator (sample mean) whose error of estimation is no greater than 0.5 with a probability of 99%.** One way to achieve this goal is to increase your sample size. Your answer in Question 1 above (based on 10 data points) gives you an estimate to the true population variance of random variables X_i's. Now calculate the minimum sample size (i.e. number of observations) needed to achieve the desired accuracy (that is, the error of estimation has to be less than 0.5 with a probability 0.99). Round your answer up as an integer. Denote the required sample size as n. Report the total cost of collecting these n observations (remember: each observation costs 12 dollars.)
5. (1pt) Use the first n data points in the data set that you have received, and use these n observations to calculate an unbiased point estimate for μ. Recall the command `x[1:n]`. Here n is the minimum

sample size that you obtained in the last question.

6. (2pts) Again, use the first n data points, provide a 90% confidence interval for μ. Note that with $n>10$ observations in your hand (which you have paid for $\$$12 each), you can get a more accurate estimate of the population variance. Remember to use the n data points to calculate a new sample mean and a new standard error. Report the 90% two-sided confidence interval, the lower bond in (6a) and the upper bound in (6b).

Answer key

```
setwd("C:/448wd")
dat = read.csv('data_3.txt',header=FALSE)
dat <- as.matrix(dat)
x <- dat[1,]

pilot = x[1:10]

ans1 = var(pilot)

ans2a = mean(pilot) - qnorm(0.95) * sqrt(ans1/10)
ans2b = mean(pilot) + qnorm(0.95) * sqrt(ans1/10)

ans3a = mean(pilot) - qt(0.95,9) * sqrt(ans1/10)
ans3b = mean(pilot) + qt(0.95,9) * sqrt(ans1/10)

nsize = ceiling( ( qnorm(0.995)/0.5*sqrt(ans1) )^2 )
# note: ceiling takes a single numeric argument x and returns a numeric vector
#   containing the smallest integers not less than the corresponding elements of x.
ans4 <- nsize*12

newdata = x[1:nsize]
ans5 = mean(newdata)
ans6a = mean(newdata) + qnorm(0.95) * sd(newdata)/sqrt(nsize)
ans6b = mean(newdata) + qnorm(0.95) * sd(newdata)/sqrt(nsize)

print( c(ans1,ans2a,ans2b,ans3a,ans3b,ans4,ans5,ans6a,ans6b) )
```

Homework 2

Round to at least 3 decimal places unless otherwise stated.

Each of you receive 225 numbers, denoted as $X_1,\dots,X_{n}$, where $n=225$. It is known that $X_i\sim Unif(0,\theta)$ independently with θ unknown.

1. (1pt) report the sample mean of these 225 numbers.
2. (1pt) correct the sample mean above so that it becomes unbiased for the purpose of estimating θ.
3. (2pts) derive the standard error of the unbiased estimator above (not the sample mean in Question 1, but the corrected one in Question 2!!!) as a function of θ, and then report the “2-standard-error bound” on the error of estimation by replacing the unknown θ in the standard error by the unbiased estimate obtained in Question 2. Round to 5 decimal places for this question.
4. (1pt) An alternative way to approximate the standard error is to estimate the (population) standard derivation of X_1 directly by using the sample standard derivation based on the data, then the standard error can be

approximated by the sample standard deviation, divided by square root n. Please report the “2-standard-error bound” on the error of estimation obtained in this way. Hint: You may use `sd(x)` to find the sample standard deviation, but you may want to use `sqrt(sum( (x-mean(x) )^2)/(length(x)-1))` to help you familiarize with the calculation (they should give you the same answer). Moreover, the “2-standard-error bound” you find here should be reasonably close to that in the last question. Round to 5 decimal places for this question.

5. (1pt) report the maximum of the 225 numbers.
6. (2pt) correct the maximum of the 225 numbers so that it becomes an unbiased estimator for θ, then report the observed value of this unbiased estimator based on the given data.
7. (1pt + 1pt) use the pivotal method to find a 95% confidence interval for θ. The pivotal quantity is transformed from the maximum of the 225 numbers. See Ex. 8.43. Then report the confidence lower and upper limits as the answers to (7a) and (7b) in the Google Form. Round to 5 decimal places for this question.

Answer key

```
setwd("C:/448wd")
dat = read.csv('data_2.txt',header=FALSE)
dat <- as.matrix(dat)
x <- dat[1,]

ans1 = mean(x)
ans2 = 2*mean(x)
ans3 = 2*2*mean(x)/sqrt(12)/sqrt(length(x))
ans4 = 2*sd(x)/sqrt(length(x))
ans5 = max(x)
ans6 = max(x)*(length(x)+1)/length(x)
ans7a = max(x)/(0.975^(1/length(x)))
ans7b = max(x)/(0.025^(1/length(x)))

print( c(ans1,ans2,ans3,ans4,ans5,ans6,ans7a,ans7b) )
```

Homework 1

Each of you are given a different data set of 64 observations drawn from an unknown distribution. Please submit your answers to
https://docs.google.com/a/binghamton.edu/forms/d/16jhF5eUpPY7pXXmzaAHAgiFZY7S8JypBYKcEfSUcBjE/viewform

Note that you need to login your Bmail account to submit the answers. Please do this by 7 pm on Feb. 2.

1. Find an unbiased estimate of the population mean based on the data set you are given.
2. Find the maximum of the observations that you receive.
3. Find the minimum of the observations that you receive.
4. Suppose we are interested in the population proportion of those observations which are greater than 4. Find an unbiased estimate of this population proportion.

The following R code may be able to help you get started. Copy each line to the console of R and press “enter”.

```
##### Assume that you have a Windows machine. First create a folder called "448wd" under C drive.
##### I trust that you can do this on your own. If not, search a solution on Google or Youtube.
##### Set the R working directory
setwd("C:/448wd")

### Read the data file. Make sure that your data file has been copied to the folder.
```

```
dat = read.csv('data_1.txt',header=FALSE)

### The variable named "dat" that you just read into R is a data frame.
### We need to convert it to a matrix

dat <- as.matrix(dat) ##dat now is a 1x64 matrix (1 row and 64 columns)

x <- dat[1,]  ## Take the first row of this matrix as your sample

# Try the following.
mean(x)  # sample mean
median(x) # sample median
max(x)  # maximum
min(x)  # minimum
y = (x > 2)
y   # We can see that Y is a logical vector of TRUE and FALSE.
### We can operate directly on a logical vector with the convention that TRUE = 1 and FALSE = 0. For
example
mean(y)
sum(y)

## Ok. You are ready to answer the questions.
```

Answer key

```
setwd("C:/448wd")
dat = read.csv('data_1.txt',header=FALSE)
dat <- as.matrix(dat)
x <- dat[1,]
ans1 <- mean(x)
ans2 <- max(x)
ans3 <- min(x)
ans4 <- mean( x > 4 )
print( c(ans1,ans2,ans3,ans4) )
```