MATH 6359, Statistical Computing, Homework 2

Andrey Skripnikov

September 12, 2017

SUBMISSION GUIDELINES:

- **Bring the hard-copy of typed up solutions to the class on Tuesday, September 26.**
- **Try to keep it under 5 pages total (all included - text, code, plots, tables). DON'T (!) repeat the problem formulation, go straight to solution.**
- **For all three problems follow the example format in terms of conciseness.**
- **For problem 2 - simply provide the code with raw output.**
- **Point total is 60 (100%), and on top of that one can get 5 extra credit points total (as a grade cushion for later in the course).**

PROBLEM #1 - 25 points.

Select a probability distribution (whichever is easily available in R, e.g. normal, binomial, uniform, Poisson, etc) and write a function that takes as input

1. **number n of values to generate from that theoretical distribution;**
2. **the values for parameters of theoretical distribution (e.g. values of μ and σ^2 for normal, or λ for Poisson)**

and does the following

1. **Generates a sample vector of n random values from your theoretical distribution with specified parameter values.**
2. **For that sample, calculates the mean, variance, median, and 99% quantile.**
3. **Outputs the absolute differences between:**

- **the calculated sample mean and the mean of theoretical distribution**
- **sample variance and theoretical variance**
- **sample median and theoretical median (which is a certain quantile)**
- **sample 99% quantile with theoretical 99% quantile**

and also a:

- **probability of witnessing THAT value of sample median OR LESS, under your specified distribution (e.g. for Binomial distribution, if the function is called for $Bin(10, 0.3)$, then output $P(X \leq sample.median \mid X \sim Bin(10, 0.3))$**

Run that function for increasing $n = 10^2, 10^4, 10^6$, comment on values you see with that increase in number of generated values (or "sample size" as you may also call it) - do the differences become larger/smaller or don't change much? What about the probability of seeing a sample median value $\leq$ than what we observed?

Example:

```
> # If you select to work with Poisson distribution.
> gener.pois <- function(n,      # Number of values to generate.
                         lambda # The parameters of your distribution.
                        ){
# Generate the sample of n values from Pois(lambda).
# Calculate the mean/median etc of that sample.
# Calculate all needed theoretical quantities (quantiles, variance, etc).
# Calculate the differences between sample and theoretical values.
# Calculate the probability of <= sample median.
# Output the vector of all those differences and that last probability.
}
>
> gener.pois(100,5)  # This should work on 100 generated values. ~Pois(5)
    mean.dif     var.dif  median.dif Q99.dif.99%    prob.val  # Type of output I expect
  0.1600000   0.1862626   1.0000000   0.9800000   0.4404933   # from your function.

> for (n in c(10^2,10^4,10^6)) print(gener.pois(n,5))
   mean.dif     var.dif  median.dif Q99.dif.99%    prob.val
  0.0600000   1.7107071   0.0000000   1.9900000   0.4404933
   mean.dif     var.dif  median.dif Q99.dif.99%    prob.val
  0.0321000   0.1455841   0.0000000   0.0000000   0.6159607
   mean.dif     var.dif  median.dif Q99.dif.99%    prob.val
0.001280000 0.001099363 0.000000000 0.000000000 0.440493285
```

PROBLEM #2 (simply provide code and raw output) - 15 points.

Write your own one-sample t-test function JUST FOR THE " $\neq$ " ALTERNATIVE (don't implement the ">" and "<" one-sided tests), name it e.g. *my.t.test()*. **It should execute steps of the testing procedure I outlined in the class. Compare your output to the R function** *t.test()* **output for the same input.**

Make sure your function takes in just those three parameters as input:

1. **data vector of quantitative values,**
2. **hypothesized mean value,**
3. **confidence level,**

and returns a list with following elements:

- **t** (**TS value**, has to equal the "$t =$ " from the *t.test()* call output for same data)
- **df** (**degrees of freedom**, has to equal the "$df =$ " from the *t.test()* call output)
- **p.value** (has to equal the "*p-value* $=$ " from the *t.test()* call output for same data)
- **CI** (has to be a 2-element vector: first element = first value under "95 percent CI" of *t.test()* output, second element = second value)
- **sample.estimates** (just the sample mean, has to equal to the value under 'mean of x' in *t.test()* output)

Some code to start you off and give you an idea of what kind of output I expect from the function:

```
> my.t.test <- function(vec,mu,ci.lvl=0.95){
... # Here the floor is yours.
}
> my.t.test(c(1:10),mu=5) # Exemplary call for your function.
$t
[1] 0.522233
$df
[1] 9
$p.value
[1] 0.6141173
$CI
[1] 3.334149 7.665851
$sample.estimates
[1] 5.5
>  t.test(c(1:10),mu=5) # Call to R's t.test() function to compare with our output.
```

```
One Sample t-test

data:  c(1:10)
t = 0.52223, df = 9, p-value = 0.6141
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 3.334149 7.665851
sample estimates:
mean of x
      5.5
>
>
> my.t.test(c(1:10),mu=5,ci.lvl=0.99) # Checking if "ci.lvl" parameter also works.
$t
[1] 0.522233
$df
[1] 9
$p.value
[1] 0.6141173
$CI
[1] 2.388519 8.611481
$sample.estimates
[1] 5.5
>
> t.test(c(1:10),mu=5,conf.level=0.99) # Compare output with similar t.test() call.
One Sample t-test

data:  c(1:10)
t = 0.52223, df = 9, p-value = 0.6141
alternative hypothesis: true mean is not equal to 5
99 percent confidence interval:
 2.388519 8.611481
sample estimates:
mean of x
      5.5
```

PROBLEM #3 - 20 points (+5 EXTRA)

Find data set that has two distinct independent groups with quantitative measurements for the subjects of those groups (it CAN'T be the ones we've already covered, e.g. *juul*, *bp.obese*, *energy*, *intake*; **check** `https://vincentarelbundock.github.io/Rdatasets/datasets.html` **for more data). For that data set:**

1. **Produce side-by-side** *frequency* **histograms for two groups, side-by-side** Q-Q **plots, and a boxplot.**
2. **Conduct two-sample** t**-test comparing the means of those groups, summarize the results.**
3. **(EXTRA CREDIT, +5 POINTS) Conduct two-sample Wilcoxon non-parametric test, summarize the results and compare those to** t**-test results.**

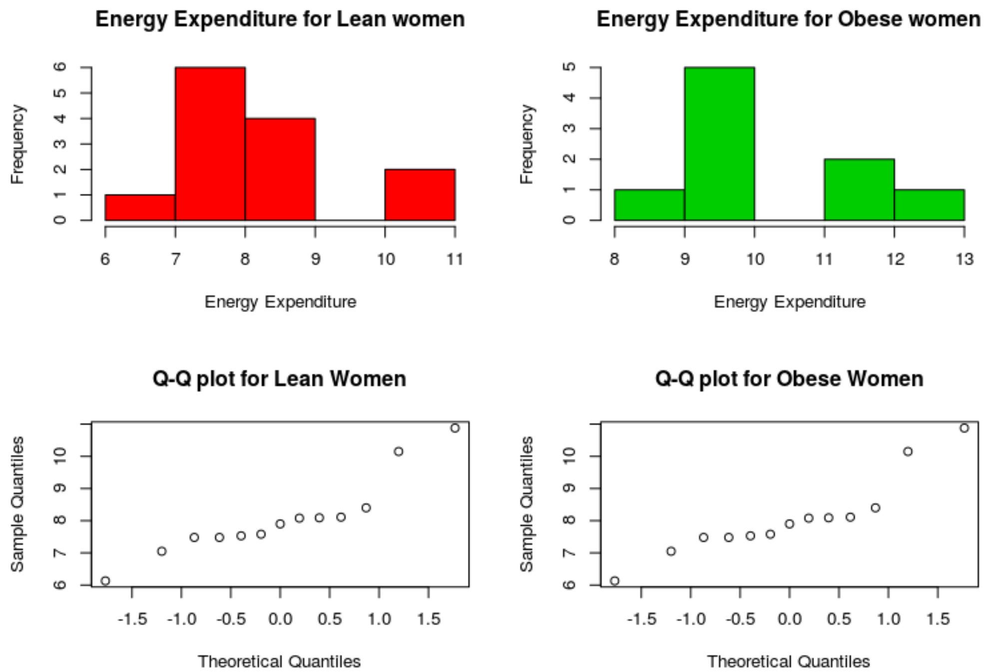
Example: I want to compare the energy intakes for lean and obese women in data set *intake* (this is just for illustration, while you guys can't use the data sets we've covered in class).

Code:

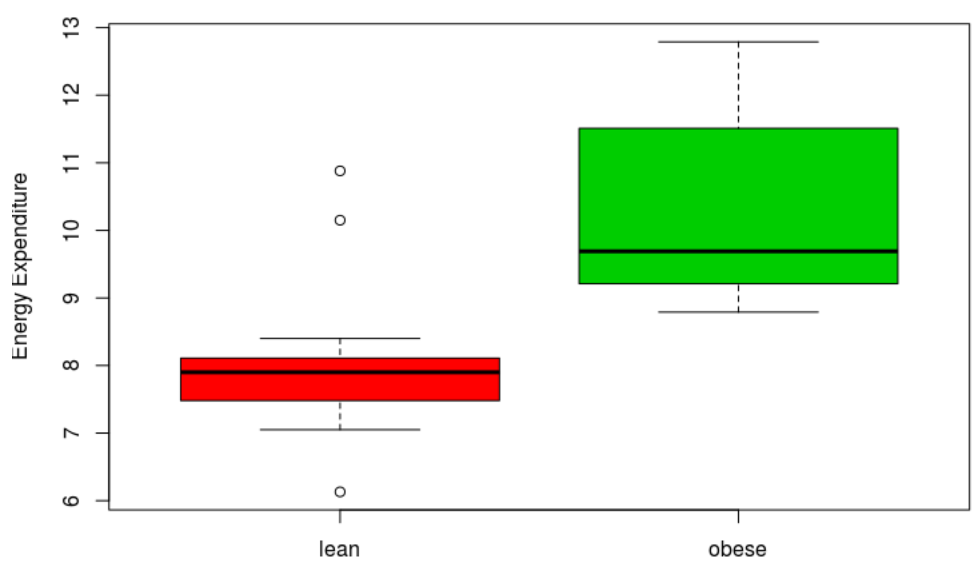
```
> # Here goes the code for the WHOLE PROBLEM:
> # Do the plots here
> hist(..)  # etc
> ###
> ## Do the tests here
> t.test(expend~stature, data=energy)  # just show me that executive command,
>  # DON'T show me the output here in the code
>
```

Plots:

I expect the histograms and Q-Q plots to look something like this



I expect the boxplot to look something like:



T-test setup and results:

Let μ_1 denote the **population mean** of energy expenditure for **lean** women, μ_2 - for **obese** women. We are looking to test if there is a significant difference in energy expenditure between women of the two statures. For that we set up the following hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0; \quad vs \quad H_a : \mu_1 - \mu_2 \neq 0$$

Preferably summarize the test results in a nice table (but raw R output will do it for me too):

Test Stat	DF	p-value	95% CI
-3.855	15.9	0.001	(-3.46,-1.00)

Summary (1-2 sentences on plots, 1-2 sentences on tests, no more):

The histogram and Q-Q plots show slight departures from normality, therefore parametric t-test results should be treated with caution. Boxplot clearly points to lower energy expenditure values for lean women than for those of obese women. The conducted t-test reinforces that belief by showing significant difference between the groups, and a fully negative 95% confidence interval for difference in expenditure between lean and obese women (meaning $\mu_1 - \mu_2 < 0 \implies \mu_1 < \mu_2$).