IST-772 Quantitative Reasoning in Data science

Week5/HW-5: Statistical Inference Part-2

Bayesian and Traditional Hypothesis Testing (Page 86-87: Problems 6-10)

Sharat Sripada (vssripad@syr.edu)

1. Run a t-test on PlantGrowth dataset for groups ctrl and trt1. Determine the t-value, p-value and confidence interval. Also using an alpha=0.05 explain if we accept or reject the NULL Hypothesis.

This exercise requires us to run a Hypothesis test on groups in the PlantGrowth dataset.

The goal is to devise a statistical test (like t-test) and conduct a Null Hypothesis Significance Test (BHST) - accept or reject the NULL Hypothesis based on the findings.

We will lay out the following steps:

1. Define a NULL Hypothesis:
   The difference in sample means of groups ctrl and trt1 is zero
2. alpha = 0.05
3. Run a statistical test like a t-test:

```
pg <- PlantGrowth

summary(pg)

ctrl <- pg$weight[pg$group == 'ctrl']
trt1 <- pg$weight[pg$group == 'trt1']

# Run a t-test between ctrl and trt1 groups of PlantGrowth data-set
t.test(ctrl, trt1)
```

```
> t.test(ctrl, trt1)

        Welch Two Sample t-test

data:  ctrl and trt1
t = 1.1913, df = 16.524, p-value = 0.2504
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2875162  1.0295162
sample estimates:
mean of x mean of y
    5.032     4.661
```

Interpreting the t-test output and Conclusion: From the output we gather the required parameters:

- t-value: 1.1913
- Degrees of Freedom (DF): 16.52
- p-value: 0.2504
- Confidence interval: -0.287, 1.829

Conclusion: Since the p-value (0.25) > alpha (0.05) we fail to reject the NULL Hypothesis.

2. Run a MCMC test between the control group (ctrl) and the treatment group 1 (trt1) in the PlantGrowth dataset and interpret the plot.

Importantly state the boundary values and explain HDI.

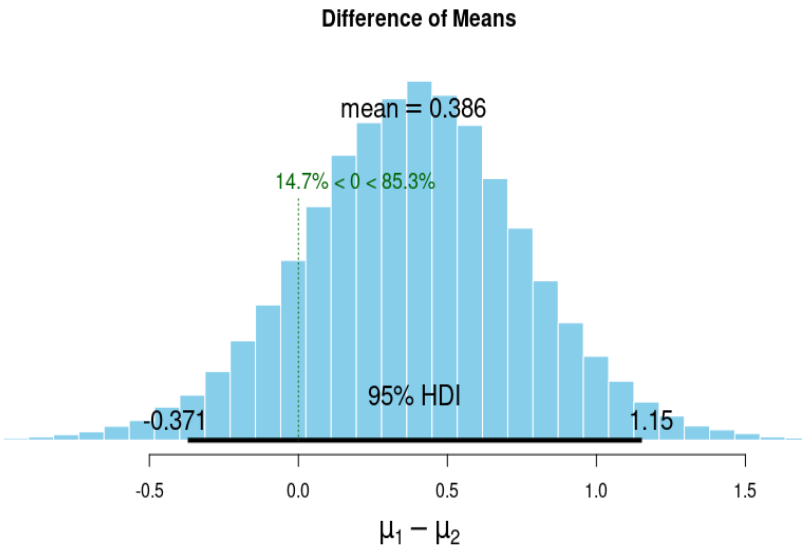
Run the R-code below to simulate a Markov Chain Monte Carlo (MCMC) test using Bayesian methodology.

Particularly the library BEST and BESTmcmc() offers this functionality:

```
# Run a Bayesian BEST simulation that uses MCMC model
```

```
install.packages('BEST')
library(BEST)
pg <- PlantGrowth
ctrl <- pg$weight[pg$group == 'ctrl']
trt1 <- pg$weight[pg$group == 'trt1']
plantBest <- BESTmcmc(ctrl, trt1)
plot(plantBest)
```

A plot of simulation results is below:



Interpreting the plot of BESTmcmc():

- The boundary values are: -0.37, +1.15
- HDI or Highest Density Interval says that there is 95% probability that the population mean difference between the two groups (ctrl and trt1) falls within the boundary values of -0.37, +1.15

3. Compare the results t-test and Bayesian MCMC test

The t-test derives a confidence interval of -0.287, 1.829 and *FAILS to reject the NULL HYPOTHESIS* since p-value > alpha.

Further, the confidence interval (CI) from the t-test can be interpreted as: 95% of all tests will contain the true population mean difference in the range of CI that is, -0.287, 1.829

In contrast, the Bayesian MCMC test interpretation is straight-forward: There is 95% probability the true population mean difference lies in the range -0.37, 1.15

Both the confidence interval from the t-test and HDI are comparable.

4. Run a t-test, Bayesian MCMC test between groups ctrl and trt2 in the PlantGrowth dataset. Like earlier document, interpret and compare interpretations.

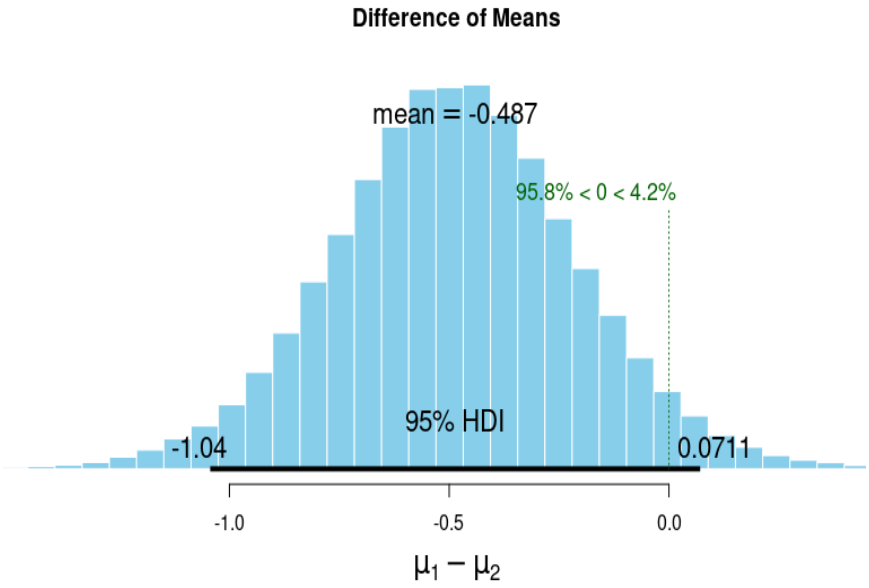
Use the following R-code to run a t-test between ctrl and trt2:

```
> t.test(ctrl, trt2)

        Welch Two Sample t-test

data:  ctrl and trt2
t = -2.134, df = 16.786, p-value = 0.0479
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.98287213 -0.00512787
sample estimates:
mean of x mean of y
    5.032     5.526
```

First and foremost, the NULL Hypothesis shall be defined as:
The difference in sample means of groups ctrl and trt2 is zero

From the t-test results or output we see that confidence interval is between -0.982, -0.005. Also, based on the p-value (0.0479) and alpha = 0.05, we would ***REJECT the NULL HYPOTHESIS***.

Finally, the BEST MCMC simulation shows the following plot:



The HDI from the MCMC test shows a range between -1.04, 0.07. This means there is 95% probability the true population mean difference lies in this range.

5. Run a t-test, on the following R-code using rnorm:

t.test(rnorm(100000, mean=17.1, sd=3.8), rnorm(100000, mean=17.2, sd=3.8))

Comment on the results. Explain the implications of using NHST on large datasets.

The t-test of a large dataset simulated using a random normal distribution with mean 17.1mpg and 17.2mpg shows the following output:

```
> t.test(rnorm(100000, mean=17.1, sd=3.8), rnorm(100000, mean=17.2, sd=3.8))

        Welch Two Sample t-test

data:  rnorm(1e+05, mean = 17.1, sd = 3.8) and rnorm(1e+05, mean = 17.2, sd = 3.8)
t = -6.5778, df = 2e+05, p-value = 4.788e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.14520057 -0.07853421
sample estimates:
mean of x mean of y
 17.08789  17.19976
```

The output shows a very small p-value (<< alpha), a statistically significant result that REJECTS the NULL HYPOTHESIS.

For large datasets, typically more than 1000 observations (situation common with big data) almost every difference, no matter how trivial is statistically significant. *And so, NHST for large datasets may not always provide useful information about the outcomes of a piece of research.*