**IST-772 Quantitative Reasoning in Data science**

Week4/HW-4: Statistical Inference Part-1

**Logic of inference using Confidence Intervals (Page-66: Problems:7-10)**

Sharat Sripada ([vssripad@syr.edu](mailto:vssripad@syr.edu))

1. **Using the built-in dataset PlantGrowth, run the summary and hist commands and interpret the results**

The summary output is as below:

> summary(pg)

weight group

Min. :3.590 ctrl:10

1st Qu.:4.550 trt1:10

Median :5.155 trt2:10

Mean :5.073

3rd Qu.:5.530

Max. :6.310

> str(pg)

'data.frame': 30 obs. of 2 variables:

$ weight: num 4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...

$ group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...

*Command interpretation:* The summary output shows two columns, weight and group – let us use str() to better understand the data in the two columns:

* Weight is type numeric and so the summary for this column shows:
  + range – min/max values
  + central tendencies – mean/median
  + quartiles – 1st/3rd
* Group is type factor and summarizes the 3 known groups – crtl, trt1 and trt2, with counts

Using the following R code, we can obtain the histogram of the weight for group ctrl:

> pg$weight[pg$group == 'ctrl']

[1] 4.17 5.58 5.18 6.11 4.50 4.61 5.17 4.53 5.33 5.14

> hist(pg$weight[pg$group == 'ctrl'])

Chart, histogram

Description automatically generated

Correspondingly, below are histograms for trt1 and trt2 groups:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

1. **Create a boxplot of the plant growth data**

Using the following R-code, we can create a boxplot:

> boxplot(pg$weight ~ pg$group)

Chart, box and whisker chart

Description automatically generated

Interpreting the box plot:

* group *trt2* has higher min, max, 1st and 3rd quartile and median values compared to groups ctrl and trt1
* group *trt1* has the lowest min values but a comparable max value with groups ctrl and trt2 and so has a larger range
* medians for all three groups are in the range 4.5 – 5.5 - trt1 is on the lower end and trt2 is on the higher end of that range

1. **Run a t-test between ctrl and trt1**

Use the following R-code to run a t-test between ctrl and trt1:

> t.test(pg$weight[pg$group == 'ctrl'], pg$weight[pg$group == 'trt1'])

Welch Two Sample t-test

data: pg$weight[pg$group == "ctrl"] and pg$weight[pg$group == "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

The Welch two sample t-test results are seen in the output above. The 95% confidence interval is shown as a range -0.287, +1.029

Interpreting the confidence interval:

* Span -0.287, +1.029 is the interval estimate of the population mean difference
* We can say that 95% of the intervals contain the true population mean difference, somewhere between -0.287, +1.029 (likely 0.371 +/- 0.658)

1. **Run a t-test between ctrl and trt2**

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

> t.test(pg$weight[pg$group == 'ctrl'], pg$weight[pg$group == 'trt2'])

Welch Two Sample t-test

data: pg$weight[pg$group == "ctrl"] and pg$weight[pg$group == "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

95% confidence interval is between -0.982, -0.005. Some take aways:

* The fact that the mean difference between ctrl and trt2 is negative tends to show that means and hence the values of ctrl are lower than trt2
* Like earlier, the confidence interval is saying that 95% of the intervals will comprise the true population mean difference