Module 0 Assignment on Reviews

Jingwen Zhong // Graduate Student

02/10/2021

Instruction

In Module 0, we briefly reviewed seven **statistical tests methods** in data analysis:

1. Conventional (z- or t-)
2. Permutation (randomization-based, all permutations)
3. Randomization or Simulated Permutation (simulation-based, some permutations, without replacement)
4. Bootstrapping (resampling-based, with replacement)
5. Linear Regression (theory-based, LS-based)
6. and 7. Nonparametric Approaches: Wilcoxon and Rank-based (median test would be used as well)

In the assignment, you will run these tests with a new data set for two-sample independent mean problem below.

An experiment was administered whether or not extra nitrogen affects the stem weight on seedlings. One group (control, n=8) was controlled with **standard nitrogen**, the other group (treatment, n=8) was given **extra nitrogen**. After two weeks, the stem weights were measured in gr. Assume the populations of the samples are normally distributed with **unknown equal variances**.

The raw data is as follows:

```
control_group <- c(0.4, 0.45, 0.35, 0.27, 0.46, 0.33, 0.3, 0.43)
trtmnt_group <- c(0.49, 0.45, 0.35, 0.38, 0.48, 0.55, 0.47, 0.65)
# boxplot(control_group, trtmnt_group) round(sd(control_group),2)
# round(sd(trtmnt_group),2)
```

```
# Or, use this way, or make data frame so you can run directly the Module0 R
# Codes
datam = cbind(c(control_group, trtmnt_group), c(rep(0, 8), rep(1, 8)))
y <- datam[, 1]
group <- datam[, 2]
colnames(datam) = c("y", "group")
# View(datam) #Check the data
```

1) (*Descriptive*) Do descriptive analysis.

a. Is the study design an experimental or observational study? Justify.

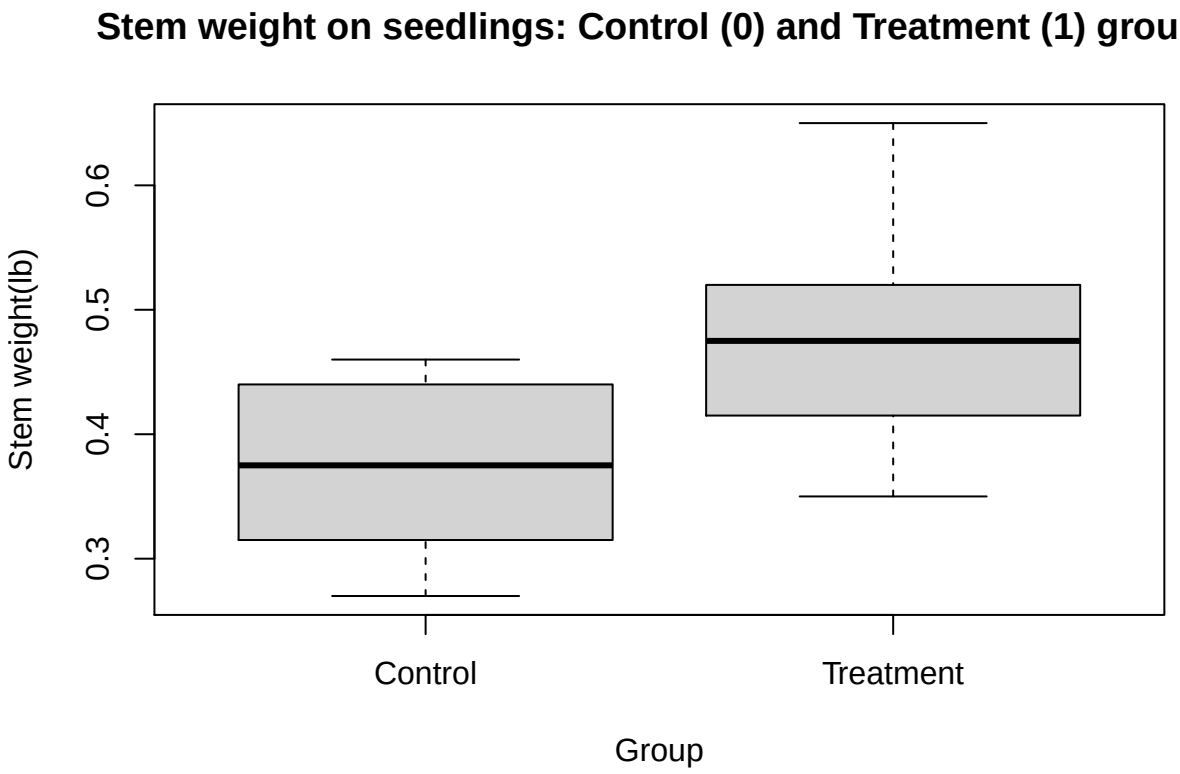
The study design is an experimental study.

In the experimental study, the researcher controls the variables to complete the study, and here the variable we control is the amount of nitrogen that being used on the stem.

The observational study relies on observation, and the researcher cannot affect the observation object

---

b. Obtain a side-by-side boxplot on measurements of the two groups. Include the graph. Add title and group names.



---

c. Obtain summary statistics that show central (mean, median etc) and spread (sd, IQR, range etc) measurements of the data distribution for each group. Make a table and include the statistics.

Table 1: Summary of the data

	size (n)	mean	sd	median	IQR
all	16	0.43	0.10	0.44	0.12
control	8	0.37	0.07	0.38	0.11
treatment	8	0.48	0.09	0.48	0.07

---

d. Compare the centers and spreads. Do you see a difference in centers? Do you see a difference in spreads? Comment.

I see a difference in centers, the center of control group is smaller than treatment group

I see a different in spreads, the interquartile range of Control group is larger than the treatment group

---

e. Do you see any potential outliers? Find and comment if exists. Explain your criterion to find outliers.

```
## z score of control group:

## [1] 0.3671239 1.0664075 0.3321597 1.4510135 1.2062642 0.6118732 1.0314433
## [8] 0.7866941

## z score of treatment group:

## [1] 0.1330885 0.2927947 1.3575028 1.0380904 0.0266177 0.7719133 0.0798531
## [8] 1.8366214
```

No, I don't see any obvious outliers, my criterion is using the Z score, and I found the z score of 65 in treatment is a little bit large, but still less than 2

2) (*Test*) Using the seven methods above, test at a 5% significance level if the difference in the mean stem weight between seedlings that receive regular nitrogen and those that receive extra nitrogen is not equal.

a. Write the hypotheses.

$$H_0 : \mu_0 - \mu_1 = 0$$
$$H_a : \mu_0 - \mu_1 \neq 0$$

*** b. What is a hypothesis testing? Write what you know about how to test in general.

Hypothesis testing is to first propose a hypothetical value for the population's parameter, and then use samples and test statistics to determine whether this hypothesis is true or not.

General steps for a hypothesis test: - set a hypotheses in terms of population parameters - Collect data and define a test statistic - Assume the null hypothesis is true, and determine if the test statistic is unlikely or not - get a conclusion

---

c. Run the seven tests and make a comparative table, showing the `p-values`. (You need to make the table showing all p-values. Modify the lab codes for this data set. Make sure you run line by line)

```
## Loading required package: e1071
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 4.0.4
```

```
## Loading required package: lattice
```

Table 2: Seven Methods Results

	P-value	Decision at 5% level
t-test	0.0262	Reject the null in favor of the alt
perm	0.0266	Reject the null in favor of the alt
perm-sim	0.032	Reject the null in favor of the alt
bootst	0.043	Reject the null in favor of the alt
lin reg	0.0262	Reject the null in favor of the alt
wilcoxon	0.0237	Reject the null in favor of the alt
rank-based	0.1053	Fail to reject the null

---

d. Conclude each result with a short comment at the specified significance level (answer part c and d together, put all the p-value calculations and the comments in a table)

see above

---

e. Now, write an `overall comment` on the results that communicate with the goal of the problem. Use the context.

The goal of the problem is to determine whether or not extra nitrogen affects the stem weight on seedlings.

The 6 results show that it does affect the stem, and we can reject our null hypothesis, while Rank-based method shows the test is inconclusive.

Overall we can reject our null hypothesis, and conclude that extra nitrogen affects the stem weight on seedlings

3) (*Concepts*) Analyze the validity of some tests.

a. List all the assumptions on `t-test`.

- Assumption of Independence: you need two independent, categorical groups that represent your independent variable.
- Assumption of normality: the dependent variable should be approximately normally distributed. The dependent variable should also be measured on a continuous scale.
- Assumption of Homogeneity of Variance: The variances of the dependent variable should be equal.

---

b. Do the assumptions meet? Check each.

Yes, it matches all assumptions

---

c. List all the assumptions on `Wilcoxon test`. Do the assumptions meet?

- Data are paired and come from the same population.
- Each pair is chosen randomly and independently[citation needed].
- The data are measured on at least an interval scale when, as is usual, within-pair differences are calculated to perform the test (though it does suffice that within-pair comparisons are on an ordinal scale).

Yes, it matches all assumptions

---

d. Do you know the assumptions on `linear regression` with LS? (a simple answer works: yes/no. if you know, write all. if not, it is fine we will learn)

NO, I don't know anything about it.

---

e. The three methods are based on randomization or resampling. What are the merits of doing randomization or resampling? Which of the three methods would work well for large data situation? Why?

Merit: In many situations, we only have one sample, and no claim about the population, we can construct a sampling distribution with randomization or resampling

Bootstrap: It is the fastest method to compute when data situation is larger. Permutation method is too expensive to compute for big sample sizes; and for randomization, the precision of estimates is usually higher, and the two sample values aren't independent

4) (*Extension*) Deep analysis and pitfalls.

a. List the `type of errors` (either Type I or II) committed in the decision made for each test. Make a table that shows all. Describe what Type I and Type II error rates are.

- Type 1 error occurs if we incorrectly reject H_0 when it is true, Type 1 error rate is false positive rate
- Type 2 error occurs if we incorrectly fail to reject H_0 when it is false, Type 2 error rate is false negative rate

Table 3: Seven Methods Results and type of errors

	P-value	Decision at 5% level	Type of error committed
t-test	0.0262	Reject the null in favor of the alt	Type 1 error
perm	0.0266	Reject the null in favor of the alt	Type 1 error
perm-sim	0.032	Reject the null in favor of the alt	Type 1 error
bootst	0.043	Reject the null in favor of the alt	Type 1 error
lin reg	0.0262	Reject the null in favor of the alt	Type 1 error
wilcoxon	0.0237	Reject the null in favor of the alt	Type 1 error
rank-based	0.1053	Fail to reject the null	Type 2 error

---

b. Build a 95% `confidence interval` on mean difference between treatment and control groups using t-critical value, df=n1+n2-2, and t-test's standard error formula. Interpret what it says. Does it confirm the p-value result from t-test?

```
## confidence interval:( 0.01364045 0.1938596 )
```

---

c. Obtain a percentile confidence interval (2.5th to 97.5th) on mean difference for the permutation test (using the permutation sampling distribution of differences on mean). Include the confidence interval. Interpret.

```
## Confidence interval is : ( -0.09375 , 0.09375 )
```

---

d. Compare the confidence intervals calculated in part b and c. Which one is more precise? Which method is more efficient? Comment.

Confidence interval in c is more precise, because sample size in t test in Part b is too small. However, t test in part b is more efficient.

---

e. Corrupt the data value .40 as 40 in the control group - as if you make a typo so this is a `bad outlier`. Recalculate the p-values of all tests. What changed? Which tests didn't change dramatically?

Only the rank-based doesn't change dramatically, the P value of other tests all changed to greater than 0.05

Table 4: Seven Methods Results

	P-value before	P-value	Decision at 5% level
t-test	0.0262	0.3445	Fail to reject the null
perm	0.0266	1	Fail to reject the null
perm-sim	0.032	0.999	Fail to reject the null
bootst	0.043	0.999	Fail to reject the null
lin reg	0.0262	0.3445	Fail to reject the null
wilcoxon	0.0237	0.1031	Fail to reject the null
rank-based	0.1053	0.2465	Fail to reject the null

---

f. (BONUS) `Standard error` is basically defined as the standard deviation of the sampling distribution of differences in mean. Either use the R packs or use the std of test statistics simulated for data. Can you find the standard error on mean difference estimate for each test? Do your best to find each. Which one(s) are most efficient test(s)? Why?

The most efficient is linear regression test

Table 5: Seven Methods Results of Standard Error

	Standard Error
t-test	0.0417
perm	0.0469
perm-sim	0.0481
bootst	0.0497
lin reg	0.0209
wilcoxon	0.0417
rank-based	0.0417

---

g. (BONUS) Ask a challenging question and answer (under the assignment context).

I hereby write and submit my solutions without violating the academic honesty and integrity. If so, I accept the consequences.

References

Retrieved 08 Feb. 2021, from https://www.statisticshowto.com/independent-samples-t-test/

Retrieved 08 Feb. 2021, from https://en.wikipedia.org/wiki/Wilcoxon_signed-rank_test#Assumptions