Applied Statistical Modelling and Inference

Simulation Assignment, due Tuesday, March 29, 5:00 pm

Instructions

In this simulation study assignment, you are expected to answer each question by conducting the appropriate analysis or simulations in `R`. You may use distributional functions such as `rnorm` as well as mechanical functions like `sum` or `sample`, but you are not allowed to use any hypothesis test already pre-coded in `R`, such as `t.test` unless otherwise instructed.

Submit your report via NYU Classes. You may either submit your report as an R Markdown document that includes both written explanations of the procedures used and results obtained as well as any code used in your analyses, or as a LaTeXtypeset document with an accompanying `.R` file. In either case, all code should be clearly organized and well-commented. We expect that you may need to consult outside resources. Please cite any resources that you use.

This assignment will be graded both on correctness of answers provided, as well as the presentation of results and the accompanying code.

Tortoise and Hare Racing Problem

After the famous fast tortoise and slow hare race, the team of 10 hares and the team of 10 tortoise had a rematch. Their finishing times are given in the dataset race.csv, in which the first column records team hares' finishing times, and the second column records team tortoise's finishing times. Your task is to follow the instructions in each of the questions to make statistical inference about the tortoise and hare race problem.

1. We are interested in testing whether the true mean finishing time is the same for team tortoise and team hare.

   (a) Specify the appropriate null and alternative hypotheses for the problem of interest using a two-sided alternative hypothesis. Comment on the implications of using a two-sided versus a one-sided alternative in this case.

   (b) Assume the finishing times of all racers are independent from each other, find the difference in sample mean in finishing times for the two teams, $\bar{X}_{hare} - \bar{X}_{tortoise}$, by calculating this quantity in `R`.

   (c) If we assume that the variance of the finishing time distributions for the two teams are equal, show that the variance of the sampling distribution of $\bar{X}_{hare} - \bar{X}_{tortoise}$ can be estimated as follows,

$$Var(\bar{X}_1 - \bar{X}_2) = \left(\frac{S^2_{hare}(N_1 - 1) + S^2_{tortoise}(N_2 - 1)}{N_1 + N_2 - 2} \right) \left(\frac{1}{N_1} + \frac{1}{N_2} \right)$$

by relating $Var(\bar{X}_1 - \bar{X}_2)$ to the quantities $Var(\bar{X}_1)$ and $Var(\bar{X}_2)$. Make sure to justify each step in your derivation. Note that S^2_{hare} and $S^2_{tortoise}$ are the sample variance of the two teams, N_1 and N_2 are the numbers of hares and tortoises. Calculate this quantity in `R`.

(d) Use the independent two-sample t-test to test the hypotheses in (a). The test statistic for such a problem given as follows,

$$t = \frac{\bar{X}_{hare} - \bar{X}_{tortoise}}{\text{std.err}(\bar{X}_{hare} - \bar{X}_{tortoise})} \tag{1}$$

i. Under the assumption that the difference in sample mean is normally distributed, this test statistic follows a t distribution with degrees of freedom $N_1 + N_2 - 2$. Calculate the test statistic and the p-value for this dataset.

ii. Setting the level of test at 5%, report the rejection region for this problem, and report your conclusion of this hypothesis test.

iii. Is the two-sample t-test used in this problem appropriate? Justify your answer by checking the assumptions of the test you just performed.

2. Let's consider a different test: if the two teams are about the same in finishing times, then we would expect the number of hares passing the number of tortoises to be roughly the same as the number of tortoise passing the number of hares. In probability terms, $P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare})$ should hold. Therefore, it is of interest to test the hypotheses:

$$\begin{aligned} H_0 &: P(X_{hare} < X_{tortoise}) = P(X_{tortoise} < X_{hare}) \\ H_1 &: P(X_{hare} < X_{tortoise}) \neq P(X_{tortoise} < X_{hare}) \end{aligned}$$

The Mann-Whitney U-test may be used to test the above hypotheses. This test is based on calculating U-statistics that look at all pair-wise comparisons between members of the two teams and summarizes the total number of wins for one of the teams.

$$\begin{aligned} U_{hare} &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_{hare,i} < X_{tortoise,j}) \\ U_{tortoise} &= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I(X_{tortoise,j} < X_{hare,i}) \end{aligned}$$

where $I(\cdot)$ is an indicator function. For example, if $X_{hare,1} = 0.5$ and $X_{tortoise,1} = 0.6$, then $I(X_{hare,1} < X_{tortoise,1}) = 1$ and $I(X_{tortoise,1} < X_{hare,1}) = 0$. Note that in order to win a race, your finishing time must be shorter than your opponent. For the sake of simplicity, assume that the times are recorded with fine-grained precision so that there are no exact ties.

(a) Calculate the U-statistic for each of the teams in `R`.

(b) Under the null hypothesis, given that there are 10 members on each team, what is the expected value of the U-statistic for each team? Explain how you arrived at this answer.

(c) Under the null hypothesis, when the sample size is large enough, the U-statistic is approximately normally distributed. The mean for this distribution, μ_{U_0}, was calculated in the previous part (2.b). The standard deviation for this normal distribution is $\sigma_{U_0} = \sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}$. Therefore, we may use the following test statistic:

$$Z = \frac{U - \mu_{U_0}}{\sigma_{U_0}} \tag{2}$$

i. Calculate the z-statistic for the Mann-Whitney U-test and report the appropriate p-value.

ii. Report your conclusion for this hypothesis test at the $\alpha = 0.05$ significance level.

iii. Mann-Whitney U test is sometimes referred to as a version of the Wilcox rank sum test. Use `wilcox.test` function in `R` to test the same hypothesis and compare your results. Set options `exact=F, correct=F` when running your `wilcox.test` function. What do these settings represent? Why are they used?

3. Permutation or randomization based tests are an alternative way to test these types of hypotheses. As with all other hypothesis tests, we must compute the sampling distribution for the test statistic under the null hypothesis. One way to construct this sampling distribution is to consider that when the null hypothesis is true, switching the group labels of the team members for the two teams should not affect the distributions of the expected outcomes or test statistics. Therefore, we can generate the null distribution by permuting the group labels for a large number of times, and computing any test statistic for each permutation. The table below illustrates several permuted datasets:

	Observed Data		Permutated sample 1		Permutated sample 2		Permutated sample 3		Permutated sample 4		...
ID\Group	One	Two	One	Two	One	Two	One	Two	One	Two	...
1	1	2	1	2	2	1	1	2	1	2	
2	1	2	2	1	1	2	2	1	2	1	
3	1	2	2	1	2	1	1	2	2	1	
4	1	2	2	1	1	2	2	1	2	1	
5	1	2	1	2	1	2	2	1	2	1	
6	1	2	1	2	2	1	1	2	1	2	
7	1	2	1	2	2	1	2	1	2	1	
8	1	2	2	1	1	2	1	2	1	2	
9	1	2	2	1	2	1	2	1	1	2	
10	1	2	2	1	2	1	2	1	1	2	

(a) Generate 3000 permuted datasets as described above.

(b) For each permuted dataset, calculate:

- $\bar{X}_{hare} - \bar{X}_{tortoise}$
- The t statistic as in equation (1)
- U_{hare} and $U_{tortoise}$
- The Z statistic as in equation (2)
- Wilcox's rank sum statistics for team Hare(W_{hare}) and for team Tortoise($W_{tortoise}$)

(c) What do you expect the mean value of each of the sampling distributions to be?

(d) For each of the quantities you calculated, use a histogram to display the sampling distribution. Comment on the similarities and/or differences you observe.

(e) For the t-statistic and z-statistic, compare the permutation-based simulated sample distributions to their theoretical distributions using Q-Q plots.

(f) For each of the quantities, test the null hypothesis by calculating simulation-based p-values. You can do this by comparing the observed test statistics with the corresponding permutation-based sampling distributions. To calculate the p-value for a two-sided test, you may assume that the sampling distribution is symmetric.

4. Summarize your findings from the first three questions by comparing your results across different tests. In which situations would you prefer one of these tests over another? Broadly comment on the pro's and con's of each of these approaches.