STA104 - Term Project

Xiaowei Zeng

2023-12-08

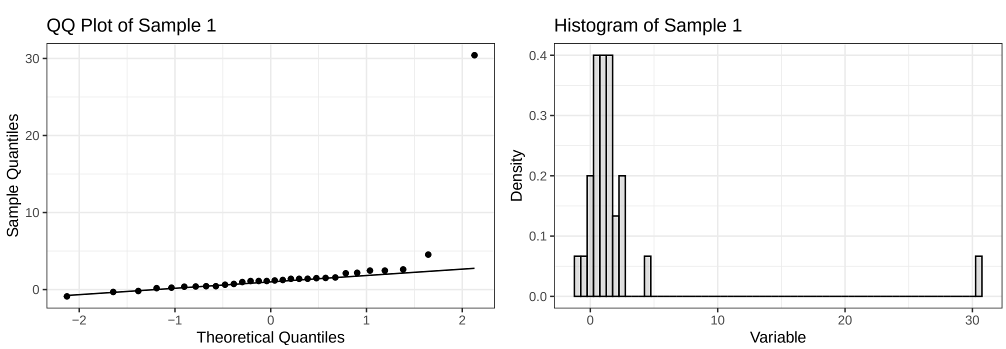
Introduction

The dataset "term project.xlsx" includes samples from several distributions. In this project, we mainly focus on carrying out reasonable hypothesis tests on different data. The significance level for all the tests in this project is 0.05. In the first problem, we conducted one-sample tests on the location shift, including t-test and binomial test. In the second problem, we conducted two-sample tests on both location shift and variance. For location shift detection, we tried t-test, permutation test and Wilcoxon Rank Sum Test. For equal variance assessment, we employed F-test, permutation F-test, permutation RMN test, Siegel-Tukey test and Ansari-Bradley test. In the third problem, we used Kolmogorov-Smirnov test to compare the distributions. In the fourth problem, we conducted four sample tests on the means, using ANOVA, permutation F-test, Fisher's Protected Least Significance Difference (LSD), Tukey's HSD and Bonferroni adjustment. Notice that the tests employed in this project are not all suitable for the corresponding data, and thus another main objective of this project is to compare the results of different tests and summarize which one may be the most appropriate for each of the samples.

Analysis of Sample 1

Description

Sample 1 has 30 observations. Before we conducted the hypothesis test, we first used QQ plot and histogram to have a quick insight of the distribution of data.



We identified a very pronounced outlier at the tail of the data, and the p-value from the Shapiro-Wilk normality test is far less than 0.001, indicating that the data is not normally distributed.

Methods

We want to carry out a test of the hypothesis $H_0 : \mu = 0$ v.s. $H_1 : \mu > 0$.

- **One-sample t-test** is a parametric test for means. It assumes that the distribution that the data sampled from is normally distributed and requires the sample size to be larger than 30. It is not robust and is very sensitive to outliers (which severely departs from the median).
- **Binomial test** is a nonparametric test for medians. No assumption need be made of the form of the population distribution. It will have greater power than t-test for heavier-tailed or higher-skewed distributions (with some extreme values and outliers).

According to the figures above, since some outliers exist and the distribution is not symmetric, which means test for medians is better than for means, it is a better choice to use binomial test than t-test.

Results and Conclusions

The p-values for both tests are shown below.

	t-test	binomial test (exact)	binomial test (normal approximation)
p-value	0.02	4.22×10^{-6}	5.89×10^{-6}

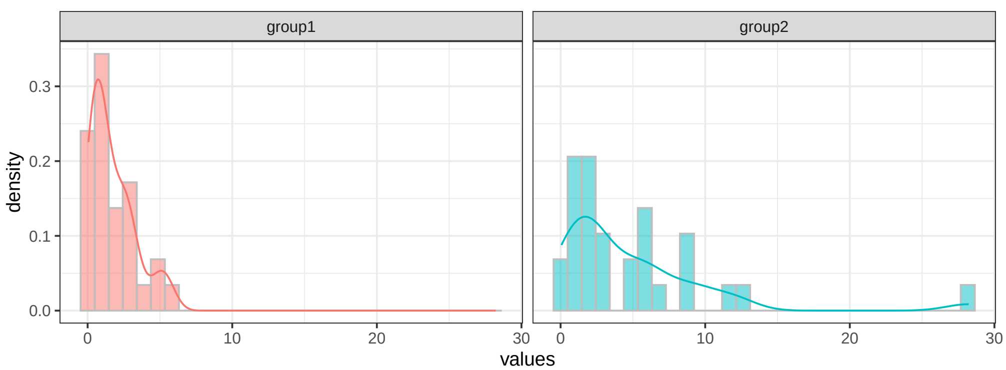
All p-values are smaller than 0.05, leading to the same conclusion that we should reject the null hypothesis, so the mean of population that the sample comes from is significantly greater than 0.

The exact p-value and the approximate p-value of the binomial test are very close and have the same order of magnitude of -6. Compared to the one-sample t-test, the p-values of binomial test are much smaller though they give the same conclusion. This can be attributed to the greater power of binomial test on distribution with extreme values.

Analysis of Sample 2

Description

Sample 2 has 2 groups of observations, each of which contains 30 data points.



The distributions of the two groups are not very similar, and some outliers exist in the group 2 data. Both of them are heavily tailed and highly skewed.

a. Test for Location Shift

We want to test the hypothesis $H_0 : \mu_1 = \mu_2$ v.s. $H_1 : \mu_1 < \mu_2$.

- **Two-sample t-test** is a parametric test for means. It has the same characteristics with the one-sample t-test introduced before.
- **Permutation test** is a nonparametric test. It requires no normality assumptions and can test for the location shift with different scores. For example, differences of means, medians, 10% trimmed means, Van Der Waerden scores and exponential scores are all employed in this problem, the power of which depend on the sensitivity of the scores to the heaviness of tail and skewness.
- **Wilcoxon rank sum test** is a rank-based nonparametric test for medians, which is equivalent to **Mann-Whitney test**. It is robust against outliers. Adjustment for ties are employed in this problem.

According to the sample distribution, we can claim that two-sample t-test which assumes normality is not suitable for this problem. We should employ robust nonparametric tests on this data.

	t-test	pt_mean	pt_med	pt_trim	pt_vdw	pt_exp	wilcox (approx)	wilcox (permt)
p-value	0.00243	0.00022	0.00246	0.00028	0.00104	0.00028	0.00141	0.00125

All p-values are smaller than 0.05, leading to the same conclusion that we should reject the null hypothesis, so group 1 is significantly shifted left from group 2.

The p-value of two-sample t-test is similar to that of permutation test for difference of medians, which are the largest among all the location shift tests. Permutation test for means, 10% trimmed means and exponential scores have similar p-values, while permutation test for van der waerden scores and wilcoxon rank sum test have similar p-values. This makes sense because the exponential score and mean score (permutation test) are very sensitive to outliers, while the rank score and van der waerden score are very robust, which means even if the outlier is large in exact value, these score functions can transform it into a relative smaller scale.

Therefore, due to the non-normality of data and the existence of outliers, the rank-based methods may be more suitable in this problem.

b. Test for Equal Variances

We want to test the hypothesis $H_0 : \sigma_1 = \sigma_2$ v.s. $H_1 : \sigma_1 \neq \sigma_2$. Since The location parameters are unknown and we cannot assume that they are the same (they are significantly different), the following tests are conducted based on the original values minus the median of their group.

- **F-test** is a parametric test assuming the normality of data.
- **Permutation F-test** and **Permutation RMD test** are nonparametric tests. The permutation F-test shares the same test statistic with F-test. The main difference between permutation F-test and permutation RMD test is that the former uses L2 norm of deviance and the latter uses L1 norm. Though both of them do not require the normality of data, but they still use the variance information of exact values. Therefore, they can also be influenced by the outliers like the F-test.
- **Siegel-Tukey test** and its extension, **Ansari-Bradley test**, are similar nonparametric tests based on ranks. Owning to the robustness of ranks, they have more power than all the other tests. Adjustment for ties are employed in this problem.

	F-test	permute F	permute RMD	Siegel-Tukey	Ansari-Bradley
p-value	4.32×10^{-10}	0.00078	0.00039	1.20×10^{-5}	1.36×10^{-5}

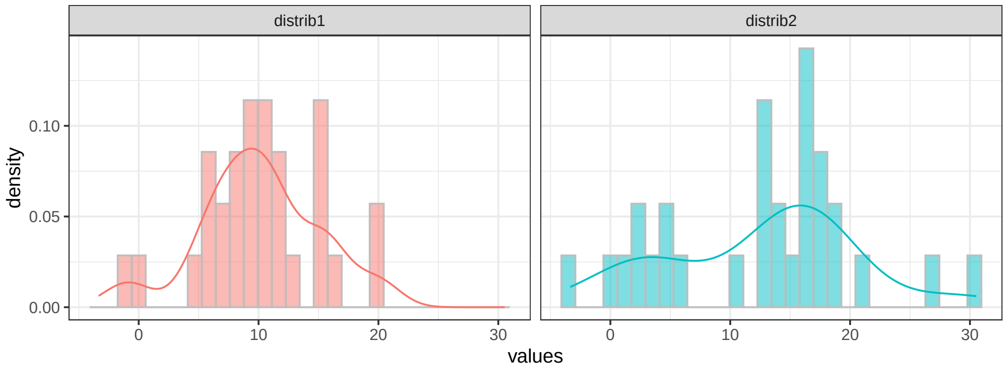
All p-values are smaller than 0.05, leading to the same conclusion that we should reject the null hypothesis, so the variance of group 1 is significantly different from group 2.

The F-test has the smallest p-value, but due to the non-normality of the data, it may not be reliable. Among the other tests, we can claim that Siegel-Tukey test and Ansari-Bradley test are the most suitable for this problem and have the greatest power.

Analysis of Sample 3

Description

Sample 3 has 2 groups of observations from corresponding distribution, each of which contains 30 data points.



According to the plots, we can find that both the mean and variance of samples from distribution 1 are smaller than distribution 2.

Methods, Results and Conclusions

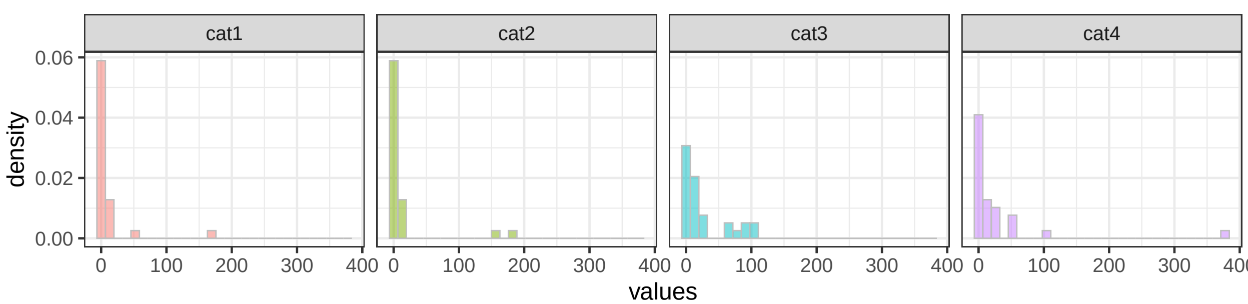
We want to carry out a test of the hypothesis $H_0 : F_1(x) = F_2(x)$ v.s. $H_1 : F_1(x) \neq F_2(x)$. However, we don't have any prior knowledge about whether the difference of distributions locates in the location parameter (mean/median) or the scale parameter (variance). Since such knowledge is not available, we use Kolmogorov-Smirnov (KS) test to compare the two distributions.

We use Monte Carlo simulation to approximate the p-value. The p-value of KS test is 0.015, smaller than 0.05, indicating that the two distributions are significantly different.

Analysis of Sample 4

Description

Sample 4 has 4 groups of observations from corresponding distribution, each of which contains 30 data points.



In each group, the data follows a heavily-tailed and highly-skewed distribution, and thus a nonparametric test is needed.

Methods

We want to test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$ v.s.H_1 : at least some means are not equal.

- **One-way ANOVA** is a parametric test for means. It is considered robust to deviations from normality and sensitive to outliers. Extreme departures from normality can impact the accuracy of the results.
- **Permutation F-test** is a nonparametric test sharing the same test statistic with one-way ANOVA which needs no assumption for normality.
- **Kruskal-Wallis test** is a nonparametric test based on ranks. It does not require the normality of data, and is robust to the outliers.
- **Fisher's Protected LSD** first conducts a Kruskal-Wallis test (or F-test). If the result is significant, multiple comparisons are then carried out. It controls the comparison-wise error at 0.05 at the cost of the explosion of overall type I error.
- **Tukey's HSD** is a nonparametric test, usually less conservative (less likely to reject H_0) than Bonferroni and more robust than LSD due to Tukey distribution.
- **Bonferroni adjustment** controls the overall experiment-wise error at 0.05 at the cost of increasing the type II error. Two-sample tests, including permutation test on means of differences and wilcoxon rank sum test, are employed here.

Results and Conclusions

	ANOVA	Permutation F-test	KW_approx	KW_permut
p-value	0.402	0.42	0.023	0.022

Though the data does not follow normal distribution, the permutation F-test has a similar p-value with ANOVA, which is larger than 0.05, indicating that there's no significant difference in any pair of groups. However, the p-value of Kruskal-Wallis test is $0.023 < 0.05$, leading to a totally different conclusion. Due to the normality of data and the existence of outliers, KW test is a better choice.

Since the result of KW test is significant, we need to do multiple comparisons between means. Because of the nonnormality of data, we should use rank-based methods in the following tests. The p-values of multiple comparisons are shown below.

X	Y	LSD_approx	LSD_pt	HSD_approx	HSD_pt	diff_pt	wilcox_approx	wilcox_pt
1	2	0.90	0.91	1.00	1.00	0.89	0.89	0.89
1	3	0.01	0.02	0.07	0.08	0.02	0.02	0.02
2	3	0.02	0.03	0.09	0.11	0.03	0.03	0.03
1	4	0.03	0.04	0.14	0.16	0.03	0.03	0.03
2	4	0.04	0.05	0.18	0.20	0.04	0.04	0.05
3	4	0.77	0.79	0.99	0.99	0.79	0.72	0.73

We first compare the p-value (normal approximation) of **LSD** with 0.05, and find that (1, 3), (2, 3), (1, 4), (2, 4) have significant difference. The permutation LSD test has similar p-values and leads to the same conclusion. However, the p-values of **HSD** (both normal approximation and permutation) give a different conclusion that no pairwise comparison is significant. For **Bonferroni** adjustment, we compare the p-values of permutation test and wilcoxon rank sum test with the adjusted α^*, 0.008, and find that no pairwise comparison is significant as well.

Therefore, LSD tells us (1, 3), (2, 3), (1, 4), (2, 4) have significant difference, while HSD and Bonferroni shows no significant difference in means for any pair of data. Because Bonferroni always make it more likely to reject H_0, and Fisher's LSD is protected by Kruskal Wallis test in the first stage, I think LSD gives a more reliable result in this problem.