STAT 788 - Midterm

Daniel Hartig

March 27, 2018

How to run

There source file for the midterm is `midterm.c`. This program takes six arguments at the command line.

Variable	Description
dist	Type of distribution for this test: 'norm' for Normal; 'normc' for Normal with contamination (sample 2 is contaminated); 'exp' for exponential.
truemean1	True mean for sample 1
truevar1	True variance for sample 1 (ignored for exponential)
truemean2	True mean for sample 2
truevar2	True variance for sample 2 (ignored for exponential)
n	Sample size (same in each group)

The main function is compiled by

```
gcc -o midterm midterm.c -lm
```

and executed by

```
./midterm exp 2 5 2 5 100.
```

In this example, the two compared distributions are normal with a mean of 2 and a variance of 5.

Methodology

When the program executes, it uses the command line arguments to determine which distributions it will compare. In the case 'norm', it will compare two normal distributions whose means and variances are as passed. In the case 'normc', it will compare two normal distributions as before, except the second sample will have a 0.02% chance of replacing each variable with a random 'large' number. Large is here defined as between five and ten standard deviations above the mean. In the case 'exp', it will compare two exponential distributions whose means are as passed. The passed variances are ignored. The mean is the scale parameter β, so the alternate rate parameterization $\lambda = 1/\beta$ is the inverse of the mean given as an argument.

The comparison of the two distributions are scored using two different methods, a two-sample t-test and a Wilcoxon rank sum test, also known as the Mann-Whitney U test. The Mann-Whitney title might be more appropriate since the implemented version uses the U test statistic.

For the t-test, the underlying distributions are assumed to be normal. The hypotheses are

$$H_0 : \mu_1 = \mu_2$$
$$H_A : \mu_1 \neq \mu_2$$

Based on the assumption of normality, a sample mean and sample variance are calculated for each sample (means as $\bar{x}_1$ and $\bar{x}_2$; variances as s_1^2 and s_2^2). The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}}.$$

The t statistic is compared against Student's t distribution to determine it's significance. Three significance levels are reported: $\alpha_{0.90}$, $\alpha_{0.95}$, and $\alpha_{0.99}$. The two-tailed critical t values for each of these significance levels are precomputed for each of the possible sample sizes ($n \in \{25, 50, 100\}$). Since the two-tailed distribution is symmetrical, a test statistic with absolute value greater than the critical value for the given sample size and significance level will lead to rejection of the null hypothesis.

For the Mann-Whitney U test, there is no assumption of underlying distribution. The hypotheses are

$$H_0 : \text{The distributions are the same}$$
$$H_A : \text{The distributions are not the same}$$

All the values from the two samples are ranked, with 1 being the lowest and $n * 2$ being the highest. The summed ranks of the first sample is calculated as R_1(either sample will produce the same result). The test statistic is

$$U_1 = R_1 - \frac{n(n+1)}{2}.$$

The test statistic is then tested against a normal distribution, which is appropriate for the sample sizes we are using. The statistic is normalized using

$$z = \frac{U - \mu_U}{\sigma_U} \quad \text{where} \quad \mu_U = \frac{n^2}{2}$$
$$\sigma_U = \sqrt{\frac{n^2(2n+1)}{12}}$$

The same three significance levels are reported: $\alpha_{0.90}$, $\alpha_{0.95}$, and $\alpha_{0.99}$. The absolute value of each test statistic is compared against the two-tailed critical value of the normal distribution for each significance level. If the statistic is greater than the critical value, the null hypothesis is rejected.

Ten thousand sample generating iterations are performed, and both tests performed for each sample. The number of rejections of the null hypothesis are recorded for each iteration. For the t-test, in the case where the means are equal and the distribution is either normal or exponential (but not normal with contamination), then the null hypothesis is true. The proportion of rejections of the null hypothesis are reported as Type I error. For all other cases, the null hypothesis is not true, and the proportion of rejections of the null hypothesis are reported as power of the test. Mann-Whitney tests that the distributions are equal, not their means are equal. Therefore, the null hypothesis is true for the Mann-Whitney test when either the distribution is normal and the means and variance are equal, or if the distributions are exponential and the means are equal.

Simulation Studies

Two normal distributions

				t-test		Mann-Whitney	
Sample 1	Sample 2	n	α	Type 1 Error	Power	Type 1 Error	Power
$\mu = 0; \sigma^2 = 1$	$\mu = 0; \sigma^2 = 1$	25	0.90	0.1001		0.0971	
			0.95	0.0523		0.0513	
			0.99	0.0099		0.0099	
		50	0.90	0.0929		0.0956	
			0.95	0.0463		0.0496	
			0.99	0.0113		0.0107	
		100	0.90	0.0984		0.0966	
			0.95	0.0495		0.0494	
			0.99	0.0102		0.0111	

				t-test		Mann-Whitney	
Sample 1	Sample 2	n	α	Type 1 Error	Power	Type 1 Error	Power
$\mu = 0; \sigma^2 = 1$	$\mu = 1; \sigma^2 = 1$	25	0.90		0.9618		0.9529
			0.95		0.9305		0.9170
			0.99		0.8014		0.7713
		50	0.90		0.9992		0.9981
			0.95		0.9973		0.9958
			0.99		0.9857		0.9817
		100	0.90		1.0000		1.0000
			0.95		1.0000		1.0000
			0.99		1.0000		1.0000
$\mu = 0; \sigma^2 = 2$	$\mu = 1; \sigma^2 = 2$	25	0.90		0.7923		0.7748
			0.95		0.6916		0.6695
			0.99		0.4409		0.4037
		50	0.90		0.9634		0.9585
			0.95		0.9327		0.9209
			0.99		0.8094		0.7853
		100	0.90		0.9939		0.9988
			0.95		0.9979		0.9977
			0.99		0.9901		0.9875

When the two distributions are identical standard normal distributions, the two tests have a similar Type I error. For both tests, the Type I error is about the same as $1 - \alpha$, which is to be expected by definition. A test with a significance level of 10% ($\alpha = 0.90$) is expected to have an estimate significantly different from the true value 10% of the time. This we observe for both tests and for all significance values and sample sizes.

When the two distributions are different, the power of the test decreases with increasing α, but increases with increasing sample size. For a large enough sample size, the power of both tests will increase to unity, as seen using the sample size 100 with means $\mu_1 = -; \mu_2 = 1$ and unit variance.

With two different distributions, the t-test is marginally more powerful than the Mann-Whitney U test for all significance values and sample sizes. However, given the variance of the sample means (σ^2/n), none of these differences are themselves statistically significant with 10,000 iterations.

Two normal distributions; the second contaminated with large values

				t-test		Mann-Whitney	
Sample 1	Sample 2	n	α	Type 1 Error	Power	Type 1 Error	Power
$\mu = 0; \sigma^2 = 1$	$\mu = 0; \sigma^2 = 1$	25	0.90		0.1098		0.1041
			0.95		0.0509		0.0516
			0.99		0.0083		0.0100
		50	0.90		0.1356		0.1089
			0.95		0.0676		0.0541
			0.99		0.0112		0.0105
		100	0.90		0.1920		0.1059
			0.95		0.1037		0.0519
			0.99		0.0202		0.0096

For two identical normal distributions, but with one contaminated with large values, both tests have difficulty distinguishing between them at $n = 25$. With this sample size, there is only a $1 - (1 - 0.02)^{25} = 0.397$ chance that any one sample even has contamination; therefore we expect about 60% of the samples to actually be from identical distributions. Therefore, the power o the test is understandably low. In fact, for the $n = 25$ case, the power of the test is basically identical to the Type I error for two identical normal distributions. This is because both tests rely on the same metric; that is, rejection of the null hypothesis. What counts as a failure for two identical distributions will instead be a success when one distribution is contaminated by large numbers (that might not even be in a small sample).

As the sample size increases, there is a 64% chance at $n = 50$ and an 87% chance at $n = 100$ that a sample with contain at least one contaminated value. One of the defining features of the Mann-Whitney U test is its insensitivity to outliers. This

is seen clearly as the power of this test does not appreciably change with sample size. The t test, on the other hand, has significantly increased power as the sample size increases. Since the t test compares means, the presence of just one large number can significantly skew the test statistic.

Two exponential distributions

				t-test		Mann-Whitney	
Sample 1	Sample 2	n	α	Type 1 Error	Power	Type 1 Error	Power
$\beta = 1$	$\beta = 1$	25	0.90	0.1004		0.0960	
			0.95	0.0479		0.0497	
			0.99	0.0079		0.0081	
		50	0.90	0.0971		0.0992	
			0.95	0.0459		0.0493	
			0.99	0.0112		0.0105	
		100	0.90	0.0985		0.1001	
			0.95	0.0485		0.0504	
			0.99	0.0108		0.0106	
$\beta = 1$	$\beta = 2$	25	0.90		0.7678		0.6679
			0.95		0.6381		0.5457
			0.99		0.3279		0.2813
		50	0.90		0.9630		0.9108
			0.95		0.9219		0.8432
			0.99		0.7530		0.6382
		100	0.90		0.9994		0.9935
			0.95		0.9975		0.9867
			0.99		0.9853		0.9451

For identical exponential distributions, the two tests have roughly equal Type I error, and their error is roughly consistent with the expected significance given the α value. This is consistent with all sample sizes and significance values.

As the two distributions diverge in their parameter, both tests have some power to distinguish between them. However, due to the nature of the exponential distribution, the change in mean of a distribution is largely driven by a few large value outliers. The cumulative distribution function of an exponential function is $1 - \exp -x/\beta$. For $\beta = 1$, we expect 63% of the values to be below 1; for $\beta = 1.5$ we still expect 49% of the values to be below 1. We would expect 86% and 74% respectively of the value to be below 2. For the rank-sum test, this represents the bulk of the values for both distributions that are well distributed with each other. At the upper end of the distribution, only a few outliers will distinguish the two distributions. For example, in a sample size of $n = 100$, we expect for $\beta = 1$ to have 0.6 values above 5, but with $\beta = 1.5$ we would have 3.6 values above 5. This difference of about 3 expected 'large' numbers may be sufficient for the t test to distinguish between the two distributions, but in many cases the outlier-insensitive rank-sum test will not be able to distinguish between the two.

Both tests gain in power to distinguish between two exponential distributions as sample size increases.

Course thoughts

I like this course a lot. The practical application of programming to the statistical methods I have been learning over the last 3 years is very useful for getting a deeper understanding about how the statistical methods work. I only regret that we couldn't take the same approach with all classes. If I was in charge of everything, we would learn a programming language in an Intro to Stats Masters class, and then every other class would be developing and delivering working statistical models for homework as we learned the theory in class.