The Wilcoxon test

Guillaume Filion

September 25, 2011

1 The problem

In 1980, John F. Laszlo studied the role of 2-chlorobenzoate on the eclosion time in larch bud moth. He obtained the following values (in hours):

```
> non_treated <- c(48.90, 20.98, 52.07, 25.57, 24.01, 51.50, 26.40);
> treated <- c(49.88, 44.00, 24.79, 45.93, 48.20, 26.51, 21.27);
```

Exercise 1
  Can we assume that the eclosion time is Gaussian? Can we apply the Central Limit Theorem?

2 The null hypothesis

What can we do if we really cannot assume that the variables are Gaussian and the sample size is too small to apply the Central Limit Theorem (roughly $n < 30$)? The t test is no longer a good option.
  We need to find another way to measure the differences between samples. If working with absolute differences does not work, we can try relative differences. After all, if all the values of a sample were higher than the values of the other, we would probably conclude that there is a difference, even if those numbers are close.

Exercise 2
  Together, propose a test based on relative differences. Formulate the null and the alternative hypotheses and design the test statistic.

3 Computing the test statistic

We have a test statistic U based on the ranks. We need to be able to extract the ranks from a collection of samples and compute the test statistic.

Exercise 3
  Compute the ranks of the sample $(2.6, 8.2, 1.3)$.
  Hint: Use the function `rank`.

Exercise 4

We have to choose which ranks to give in case the values are equal. What does the function `rank` do with ties by default?

Exercise 5

Write a function that computes the U statistic (of the first of two samples). Use it to compute the U statistic of the data.

4 The distribution of U

We are now faced with a problem: what distribution are we going to use to generate samples under the null hypothesis in order to compute the U statistic?

The beauty of this approach is that **it does not matter!** Indeed, if ranks are sampled from the exact same distribution, it is like picking the first 14 numbers and assigning 7 at random to sample 1 and the 7 other to sample 2. We will verify this now.

Exercise 6

Generate 10,000 values of U obtained from 2 standard Gaussian samples of size 7. Store the result in a vector `U.Gaussian`.
Hint: Use a `for` loop.

Exercise 7

Do the same with 2 exponential samples and store the result in a vector `U.exp`. Then do the same with 2 uniform samples and store the result in a vector `U.unif`. Finally, overlay the density distributions of those computations of U.
Hint: Use `rexp` and `runif`.
For the aces: what happens to the U statistic if samples differ only through their spread? Is the Wilcoxon test robust to the assumption that both distributions must have the same spread?

5 The decision rule

Now we need a decision rule. We will use the same approach as for the t test, *i.e.* define a rejection region at level 0.05, symmetric and away from the central bulk of points. The difference is that the distribution of U is not centered, so we need to know this central value.

Exercise 8

What is the exact expected value of U?

Exercise 9

Get the approximate limits of the rejection region and finish the test. Estimate the p-value and compare your results with the output of `wilcox.test`.

Answer of Exercise 1

No: the data is bimodal (eclosion time depends on the hour of the day). No: the datasets are too small.

Answer of Exercise 2

The idea is to work with the ranks of the values. The null hypothesis of the test is composed of a single statement:

1. Sampling is IID across samples.

The null hypothesis is very general, so you might think that we have found the perfect test. However, the alternative hypothesis is somewhat restrictive:

1. Sampling is IID within each sample.

2. Distributions differ by a 'horizontal' shift.

Thus, not every difference in distribution will lead to the rejection of the null hypothesis. A good test statistic to choose for this approach is the mean of the ranks of the values of the first sample. For example, if we have two samples $(1.2, 10.3)$ and $(3.4, 5.0, 17.1)$, the test statistic is $(1 + 4)/2$. We will call this statistic U.

However, the literature is not unanimous on which statistic to use, several others are available. The statistic used in R, is W. If the two samples are denoted $(x_1, ..., x_n)$ and $(y_1, ..., y_m)$, W is the number of pairs (x_i, y_j) for which $x_i \geq y_j$. In the previous example, the statistic is 2.

Answer of Exercise 3

```
> rank(c(2.6, 8.2, 1.3));

[1] 2 3 1
```

Answer of Exercise 4

It gives the average rank.

Answer of Exercise 5

```
> # Didactic version:
> U <- function(x, y) {
+     allranks <- rank(c(x,y));
+     ranksx <- allranks[1:length(x)];
+     return (mean(ranksx));
+ }
> # What I would write:
> U <- function(x,y) {
+     # Return the U statistic for sample x.
+     mean(rank(c(x,y))[1:length(x)]);
+ }
> U.obs <- U(treated, non_treated);
> U.obs;

[1] 7.428571
```

Answer of Exercise 6

```
> U.Gaussian <- rep(NA, 10000);
> for (i in 1:10000) {
+     U.Gaussian[i] <- U(rnorm(7), rnorm(7));
+ }
```

Answer of Exercise 7

```
> U.exp <- rep(NA, 10000);
> U.unif <- rep(NA, 10000);
> for (i in 1:10000) {
+     U.exp[i] <- U(rexp(7), rexp(7));
+     U.unif[i] <- U(runif(7), runif(7));
+ }
> plot(density(U.Gaussian), main="Density plots", xlab="Value");
> lines(density(U.exp), col=2);
> lines(density(U.unif), col=3);
> legend(x="topright", legend=c("Gaussian", "Exponential", "Uniform"),
+     lwd=1, col=c(1,2,3));
```
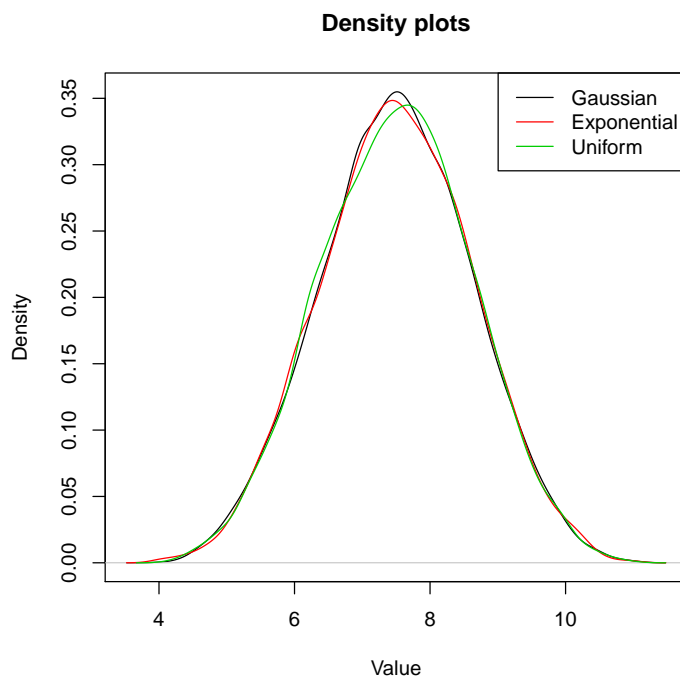

Answer of Exercise 8

The sum of the ranks of the first sample ($\tilde{U}_1$) and the second sample ($\tilde{U}_2$) is always $1 + 2 + ... + 14 = 105$. Because of symmetry, the expected values of $\tilde{U}_1$ and $\tilde{U}_2$ must be equal.

$$E\{\tilde{U}_1 + \tilde{U}_2\} = 2E\{\tilde{U}_1\} = 105, \text{ so } E\{\tilde{U}_1\} = 105/2 = 52.5$$

By definition, the statistic U is $\tilde{U}_1/7$ So $EU = 7.5$.

Answer of Exercise 9

```
> # Get an estimate of the acceptance region.
> quantile(U.Gaussian, probs=c(0.025, 0.975));

    2.5%    97.5%
5.285714 9.714286

> # The null hypothesis is accepted.
> # To check the p-value, we center and symmetrize the
> # statistic U statistic.
> mean(abs(U.Gaussian - 7.5) >= abs(U.obs - 7.5));

[1] 1

> wilcox.test(treated, non_treated);

        Wilcoxon rank sum test

data:  treated and non_treated
W = 24, p-value = 1
alternative hypothesis: true location shift is not equal to 0
```