CHAPTER 16

The difference between sports and business: Thoughts on the *t* test and the Wilcoxon test

Sports is to business as Wilcoxon is to *t* test

I am sure you can track down some post-deconstructionist who'll tell you that sports is business by other means. (Or maybe the other way around.) But here's how I look at it: in sports, rank matters; in business, amount matters. In sports you want to come first; in business, you want to make a lot of money. The Mets might have a mediocre season, ending only a few games

over .500, but if they are first in their division they'll go to the playoffs—end of story. The owners of the McAllister Towing and Transportation Company don't really care whether they are the most profitable tugboat operation in the New York area—they just want to make a profit and it will hopefully be more than a few dollars past breaking even.

As regards statistics, the Wilcoxon is like sports and the *t* test is like business. Here is how a Wilcoxon test works: you compare whether ranks are higher in one group than the other. Here is how a *t* test works: you compare whether the mean is higher in one group than the other group.

Analysis of a sports experiment

In the Tour de France, each team is accompanied by a *soigneur*, or "healer," whose job is to give massages to help the cyclists recover from each day's ride. Imagine that you are the coach of the Columbia University cycling team and that your team has picked up a volunteer *soigneur* from a local massage school. This seemed like a good idea at first, but you are now rather tired of the *soigneur* interrupting training to discuss mystical energy flow and you want to find out for sure whether massage actually helps. So you run the following experiment: on Sunday, your team takes part in a 50 mile race. That afternoon, you randomly select half of your team to get a massage. On Monday, you send everyone for a time trial. You then look at the time trial data to address the null hypothesis that the cyclists receiving the massage were no quicker than those who did not receive a massage.

Here are the results, with the times given as minutes:seconds:

Massage group	Control group
51:55.1	48:49.9
53:39.7	53:17.4
58:29.8	59:33.6
59:22.8	60:49.4
59:24.1	61:12.7
59:57.2	62:33.6
60:32.1	63:18.7
61:43.3	63:19.2
63:13.4	65:15.5
63:40.3	65:25.4

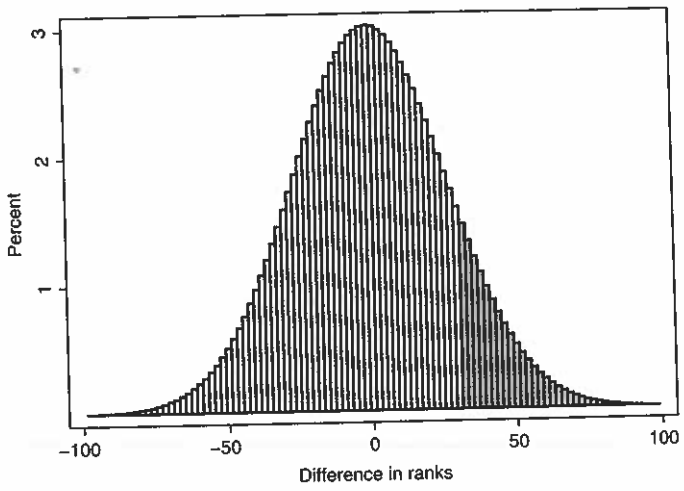
To do a *t* test with these data, you work out the mean and standard deviation in each group. You get a mean of 59.20 (SD 3.788) minutes in the massage group and 60.36 (SD 5.337) minutes in the control group. If you plug these numbers into the appropriate formula, you get a difference

between groups of 1.16 minutes. This means that, on average, cyclists receiving a massage completed the time trial about 1 minute and 10 seconds faster than those in the control group. The standard error of this difference is 2.07. We know that, if the null hypothesis were true, then 95% of the time the difference between groups would be no more than two standard errors away from zero. For this experiment, we are about half a standard error away from zero and so we know we have a non-significant result (it is actually $p = 0.6$) and can't reject the null hypothesis of no difference between groups.

To do a Wilcoxon, you have to work out the ranks of the data—who came first, who came second and so on. The cyclist coming in first was in the control group with a time near 49 minutes, and this individual gets a rank of 1. The next fastest, a full three minutes back, was a cyclist in the massage group, who gets a rank of 2. If you keep going assigning ranks in terms of where each cyclist came in the time trial, what you get is something like this, with the ranks shown in the brackets:

Massage group	Control group
51:55.1 (2)	48:49.9 (1)
53:39.7 (4)	53:17.4 (3)
58:29.8 (5)	59:33.6 (8)
59:22.8 (6)	60:49.4 (11)
59:24.1 (7)	61:12.7 (12)
59:57.2 (9)	62:33.6 (14)
60:32.1 (10)	63:18.7 (16)
61:43.3 (13)	63:19.2 (17)
63:13.4 (15)	65:15.5 (19)
63:40.3 (18)	65:25.4 (20)

If you add up the ranks in each group, you get a total of 89 in the massage group and 121 in controls, a mean of 8.9 and 12.1. So, the average cyclist receiving a massage would come in 9th—three places ahead of the average cyclist not receiving massage, who'd come in 12th. (Ok, I know—you can't get a whole bunch of cyclists all coming 9th in a time trial, but you take my point.) Our question now is whether the difference in average rank of 8.9 and 12.1 is statistically significant. To get a p-value, we can think back to the definition of the p-value in terms of the probability of the observed data or something more extreme—if the null hypothesis were true. What we'd expect if the null hypothesis were true is that there would be no difference in rank between the two groups. But we wouldn't be surprised if there was sometimes a small difference in ranks and occasionally we'd expect see a big difference, just by chance. This figure shows the difference in ranks if the null hypothesis were true and massage didn't have any effect on cycling performance:



As you can see, you get a normal distribution. I've shaded all bars where the difference in ranks was as high or higher than what we saw in our study, which was 32. You can see that it isn't particularly unusual to get a difference in ranks that large and so we'd guess that we don't have a statistically significant difference. If you add up the height of the shaded bars (the *p*-value is the probability of the observed data or *something more extreme*, meaning a bigger difference in ranks), you get 0.113. But remember, we have to take into account the possibility that we might get better results in the control group, and that this would also lead us to reject the null hypothesis. When you add up the height of the bars for a difference in ranks of −32 or less, you get, naturally, 0.113. Adding 0.113 and 0.113 gives a total probability of 0.226. So we can state that there is a 22.6% chance that you would see a difference in ranks of 32 or more if the null hypothesis were true, that is, no effect of massage on cycling times. This is the *p*-value you get when you run a Wilcoxon test on these data.

So you should fire the soigneur, right?

The answer to that question probably depends on just how annoying you found all that talk of energy chakras. The key thing from a statistical point of view is that our non-significant *p*-values don't mean that we accept the null hypothesis. The fact that the data would be reasonably likely if the null hypothesis was true doesn't mean that the null hypothesis is true, after all, the data would also be reasonably likely under a hypothesis that massage did improve cycling times, but only by 15 seconds.

The key thing to remember for now is that the cycling experiment helps explain the difference between the *t* test and the Wilcoxon. For the *t* test, we first calculated an estimate that addressed our research question: we wanted to know if massage improves cycling times and so we calculated that, on average, cyclists receiving the massage were quicker than controls by a little over a minute. We then calculated a standard error for this estimate and divided one by the

other to calculate a *p*-value. For the Wilcoxon test, we converted all the results into ranks, calculated the difference in ranks between groups and compared that difference in ranks to what we'd expect if the null hypothesis were true (that is, no difference in rank).

So what is better, Wilcoxon or *t* test?

The best thing about the result of the *t* test is that we got an estimate for the effect of massage on cycling performance: massage improved time trial times by a mean of 1 minute 10 seconds. We can also calculate a 95% confidence interval, which is the mean plus or minus about twice the standard error. The standard error was around 2 minutes and 10 seconds, so we get a confidence interval of −3 minutes 10 seconds to 5 minutes and 30 seconds. Our best guess is that massage reduces times by a minute or so, but it could actually slow you up by 3 minutes or lead to a dramatic 5 minute reduction in race time. All this is wonderful and terrific information, but only if it is correct. As Bill Gates found out when he walked into the diner (see *So Bill Gates walks into a diner: On means and medians*), calculating a mean of something only makes sense in certain circumstances. You'd hardly want a *t* test to compare the difference in salaries between Dizzy's diner (where Bill is tucking into a three egg omelet) and the Little Purity diner down the block (which has no visiting billionaires) because you'd calculate a difference in mean salary of $125m, with a 95% confidence interval from −$125m to +$375m. As you can tell from this example, the *p*-value you get from a *t* test is not very reliable when the data are very skewed—in simple terms, the *p*-value is too high if there is a true difference between groups.

Moreover, although I suggested that getting an estimate is a good thing (it normally is), estimates aren't always interesting. If a biologist has some complicated hypothesis about the effects of a gene, and conducts a mouse experiment, the estimate might be something like "IL-2 production from PMA stimulated CD45RA + CD4 + cells was increased by 0.002 units in knockout mice." It is unclear whether this has any meaningful interpretation. The main point of such laboratory experiments is to investigate hypotheses and, as such, it is only the *p*-value that is of interest.

So, again, it all depends. This makes sense, because if you could really say whether the *t* test or Wilcoxon was better, whichever one was worse wouldn't be used anymore and I wouldn't have to write chapters about it.

• Things to Remember •

1. Statistical tests are applied to data to generate *p*-values to test hypotheses.
2. The *t* test and the Wilcoxon test are two well known statistical tests.
3. Both tests are used when two groups (such as boys and girls or treatment and control) are compared with respect to a continuous variable (such as time to complete a cycle race).
4. The *t* test involves calculating an estimate addressing the hypothesis of the study as well as its standard error. The *p*-value is calculated by comparing the estimate to the standard error.
5. To run a Wilcoxon test, the data must first be converted to ranks. The *p*-value is calculated by comparing differences in ranks to an expected distribution of differences in ranks.
6. The *t* test can be unreliable if the data are very skewed.