Statistics Investigation

Planning

Introduction

I will investigate performance in the Olympic triathlon event using data gathered from the Olympic results website. The first Olympic triathlon was held at the Summer 2000 Olympics so data should be available for three Olympics (2000, 2004 and 2008). Men and women compete in separate events.

I will use a census of the individual years and genders because in Autograph I can easily manipulate large data sets (there are typically 50 competitors in each event). Autograph will enable me to compare statistics for each data set, using box and whisker diagrams showing extreme values, quartiles and the inter-quartile range. I will also create scatter graphs of one data parameter against another so that I can see if there is any obvious relationship between them.

Autograph can also easily fit trend lines (to enable predictions to be made) and calculate the Spearman's ranking coefficient. A strong correlation could indicate, for instance, that the results in one sport were very likely to influence the overall results whilst a weak correlation would indicate that an athlete could do well or badly in that sport and still be successful overall.

I am hoping to find what kind of athlete does best in a triathlon: must they be very good in all three sports, or are some sports more critical than others?

I will also investigate the differences between men's and women's times to see whether the same patterns are seen in both and whether conclusions from the men's race are equally valid for the women's competition.

First Hypothesis

The three sports will be equally important, such that there is no clear priority in terms of importance. I would expect the length of each event (1500m swim, 40 km cycling, 10 km running) to have been chosen so as not to favour any particular type of athlete.

How I hope to support this hypothesis

I will investigate this hypothesis using a quota sample of just one competition (one year, one gender – 2000, Men). This will give a sensible number of graphs (6, see below) whereas plotting equivalent graphs for all 6 data sets (men and women, years 2000, 2004, 2008) would give 36 separate graphs which would be difficult to assess. (The data should not be combined into a single data set because there may be significant differences between the data sets due to, for instance, temperature, altitude, wind or gender).

I will use the "Men, year 2000" data set: it is simply the first in the table. (I could of course repeat my study with women's data to determine whether my findings apply to men only or to both men and women).

I will plot three scatter graphs, using Autograph, to compare overall position against position in swimming, in cycling and in running. There may be an obvious correlation in some or all of these graphs.

I will rank the data in Excel so that for each athlete I know their position, both in the individual events and overall. What matters is who wins (the position) rather than the actual times and plotting data in this form will avoid any false origin and clearly indicate those who are strongest or weakest.

For each scatter graph I will use the results box to find the Spearman's rank correlation coefficient. A coefficient of 0.5 or greater will indicate a positive correlation.

<u>If the hypothesis is correct, what do I expect?</u>

A high positive correlation would indicate a sport that is particularly important to the overall result; conversely a negative correlation would indicate that it is actually best to do badly in that competition.
I expect that all three sports will show a positive correlation coefficient of at least 0.5.
I further predict that the range of correlation coefficients will be 0.3 or less (e.g between 0.5 and 0.8).

<u>Further investigation of hypothesis 1</u>

If all sports are important, I would expect the overall winners to be good in all three, and the losers to be poor in all three. This means that there will be a positive correlation <u>between</u> the sports.

Using the same data sets, I will plot scatter diagrams of:

- cycling versus swimming
- running versus swimming
- running versus cycling

These will show me something about athletes in general: is the conclusion from investigation 1 inevitable because good athletes are always strong in all 3 sports and vice-versa?

<u>If the hypothesis is correct, what do I expect?</u>

This will not disprove the hypothesis but it will provide additional insight: if some people are good at all three sports, does that mean they are exceptional or just that they would be unlikely to be good

at just one? If it is the case that they cannot be good at just one event but will always be good (or bad) at them all, then the hypothesis itself is inevitable.
I would expect to see:

- a negative correlation between swimming and cycling, since swimming uses arm muscles and cycling uses leg muscles.
- little correlation between running and swimming, since for efficient long distance running an athlete needs more springy muscles but for swimming they do not.
- some correlation between cycling and running, since they both rely on leg muscles.

This may however just tell me something about the type of athletes who choose to do a triathlon.

Predicting one result from another.

I will add a "best fit" trend line to the cycling versus swimming graph since I am expecting an obvious negative correlation. This will let me predict for another year, the cycling position from the swimming position.

Potential problems – first hypothesis

Since this is secondary data (timed by the race officials, written down, then typed into a computer) there is the possibility that there was some human error at some stage in the process. Since the Olympics is an important and carefully managed competition I presume that great efforts were made to avoid such errors: perhaps using a video recording of the finish line as an electronic record against which the written records could be cross-checked.
It is nevertheless conceivable that the data has been corrupted at a later stage, perhaps through accidental mis-typing or dragging cells in Excel.
When I create box and whisker diagrams of the time (not ranked) data for hypothesis 2 I will look at any extreme values these reveal and consider whether they are sensible data or, possibly, an indication of some anomaly – an outlier.

A point will be classed as a possible outlier if it is more than 1.5×IQR below the first quartile or more than 1.5×IQR above the third quartile. This limit is defined in the EdExcel GCSE Statistics book, p143.

If the time is unexpectedly short, that implies the athlete is vastly better than all the others. This is not likely but since the athlete's names are included in the data set I can check whether, perhaps, they set a new World Record –i.e. whether the time was recognised as a genuine value. It is possible in some sports (cycling in particular) that many competitors will ride as a group so the IQR might be very small. It is then possible for a fast competitor to be beyond the usual outlier limit without there being any kind of error. If however the IQR was *not* unusually short (instead the data value was small or large in absolute terms) one might conclude that there has been a mistake in recording the data. If this happens the validity of the data will be checked against another source. If no error can be found, it must be assumed that the value is genuine.

If the time is slightly too long, that might imply that the athlete had an injury or accident. It is still genuine data and, being probably only one point amongst the 50 competitors, would have little effect on a ranking coefficient. If it is much too long that might imply that the data was corrupted. I would then remove that athlete from the data set, saving the file with a new name so that its provenance is obvious.

Second hypothesis

The women's results in each competition will be similar to the men's apart from a shift in the times: the mean and median time may increase but the interquartile range in each competition will be very similar.

How I hope to support the second hypothesis

I will use Autograph to create box and whisker diagrams comparing the men's and women's overall performance in 2000, 2004 and 2008. Besides looking at the diagrams for a visual assessment of the trends, and to confirm there are no outliers, I will record the median and semi-IQR that Autograph calculates.

The semi-IQR will be typed into Excel (values in Appendix) and doubled in Excel before pasting back into Word.

If the hypothesis is correct, I expect to see that for each year, in each sport, the women's box is similar in width to the men's but translated sideways. I expect the difference between the men's median time and the women's median time, in each sport, to be similar from year to year. To be more specific, I expect women to be a certain percentage slower than the men, so I will calculate the ratio

$$\frac{\text{Women's overall median time}}{\text{Men's overall median time}}$$

Tip: you can write equations using the Equation Editor or, even better, MathType

for each year. I expect this ratio to remain constant even though the actual times may vary from year to year due to, perhaps, altitude or weather.

I also expect the inter-quartile range (IQR) for the men and women to be similar. Women may be slightly slower but I do not expect a wider spread of times.

Second hypothesis – further investigation

If there is an overall difference in speed between men and women, it is interesting to see whether it is larger in some sports than others.

I expect a larger difference in the swimming events (which rely on upper body strength) than in running and cycling (leg muscle events).

I will therefore create box and whisker plots for the three individual events and compare median times for men and women in each by calculating the ratio

$$\frac{\text{Women's median time}}{\text{Men's median time}}$$

in each sport. Since both men and women compete over the same distance *D* (in any one event) I can use the formula for speed in terms of distance and time $S = \frac{D}{T}$ to calculate the ratio

$$\frac{S_{men}}{S_{women}} = \frac{\left(\frac{D}{T_{men}}\right)}{\left(\frac{D}{T_{women}}\right)} = \left(\frac{D}{T_{men}}\right) \times \left(\frac{T_{women}}{D}\right) = \frac{T_{women}}{T_{men}}$$

Potential problems – second hypothesis

The numbers of men and women competing each year are not identical – the men/women split is not exactly 50/50.

Year/MW	2000-M	2000-W	2004-M	2004-W	2008-M	2008-W
Athletes	48	40	45	44	50	45

If one assumes that the winners (men or women) are the best in the world, this must imply that the shorter data set did not leave out any "better" athletes at the winning end; rather, it left out "poorer" athletes at the slow end. The corollary of this is that (in 2000, for instance, with an excess of men) the last 8 men pull the median towards a longer time than one would expect with just 40 competitors.

This is a source of bias in the median values. In an attempt to reduce the bias, I will use for my box and whisker plots a subset of each year's data containing the best *N* men and *N* women, as indicated by their overall position, *N* being the length of the smaller (men or women) data set.

Third hypothesis

I expect the distribution of times (in each sport and for both genders) to have a positive skew, since it is easier to be 10% slower than the median than 10% faster. There is a minimum time achieved (the world record, maybe) but no kind of upper time limit.

With a positive skew, the Normal distribution will probably not fit the data particularly well. Nevertheless, knowing (course book) that for a Normal distribution the probability of a random

value lying more than 2 standard deviations from the mean is 1 – 0.9772 = 2.3% (very unlikely), I predict that out of the 18 individual competitions (3 years, swimming/cycling/running, male/female) there will be no athletes whose time is more than 2 standard deviations less than the mean in their competition.

Justification

Using the binomial distribution formula and assuming the probability of an athlete finishing in a time less than 2 standard deviations below the mean is 0.023 as above, the probability of 1 instance of this rare event happening in 18 competitions will be:

$^{18}C_1 \times 0.023^1 \times 0.977^{17} = 0.28$ This is less than 0.5 so is "unlikely".

Alternatively, the probability of this event not happening at all in 18 competitions is

$0.977^{18} = 0.66$ so the probability of an athlete breaking this limit one or more times in 18 competitions will be 1 – 0.66 = 0.34 (still unlikely)

Method:

I will plot a histogram showing the distribution of times for three of these data sets and compare it (visually) with a Normal distribution.

I will use an unequal interval histogram since there are probably many athletes finishing within a short time period in the middle of the race, but fewer arriving first or last. Autograph offers the choice between "frequency" or "frequency density" as the y-axis parameter and I will choose frequency density since (with unequal intervals) this gives a good visual indication of the frequencies (= area of bars) that is easy to compare with the Normal curve.

The Normal curve in Autograph is defined as a probability distribution such that the total area under the curve = 1. Autograph can automatically fit a Normal distribution to the data and scale it up to be a frequency density distribution but this does not let one manually control the Normal curve parameters.

I will, as an alternative to this, divide all the group frequency densities by the total frequency so that each is a relative frequency density (= estimate of probability density, in the same way that relative frequency is an estimate of probability). This is very easy to do since in Autograph's "Histogram Options" dialogue box I can enter a unit of, for instance, 1/40 if there are 40 athletes in the competition.

I will then in Excel tabulate the mean, standard deviation and minimum time for all these events to see if any are in this category.

Potential problems

It may be that the distributions are so far from the Normal shape that the 2.3% point is not 2 standard deviations from the mean. This could be because they are skewed or because the peak is

either flatter or pointed than in the Normal curve. If the data is skewed because of just a few people with very long times I will try fitting a Normal curve “by eye” to fit most of the histogram and ignore the right hand tail.

Data collection, processing, calculations and plotting

(a) Data collection and Excel processing.

The data was downloaded from http://nrich.maths.org/content/id/8061/TriathlonResults.xls

This Excel spreadsheet was saved onto my hard disk and a copy made to allow checking in case errors in processing were thought to have corrupted the working version. I set the copy to be read-only to avoid any risk of editing the wrong one by mistake.

The timings were originally listed in hour:minute:second format. These were converted into seconds using (e.g. for the first data cell in each sheet) the command =C3 * 86400. This technique was found by a Google search for "convert Excel time to seconds".

Having created one such cell, it was dragged sideways, then downwards to copy the calculation for all the data on that sheet. The cell format was set as number with 0 decimal places as it was felt that fractions of a second would be unnecessarily accurate for the task in hand. (The rounding did not affect Excel's ranking calculation which would be based on the exact cell values, not the displayed version).

Ranking was done using the command (for instance) =rank.avg(h3, h$3:h$53, 1). The "$" here allows the command to be duplicated whilst referring to a fixed overall list.

To identify possible outliers, Excel was used to calculate, for each event:

- minimum time (using the =MIN(range) function)
- first quartile Q_1 (using the = QUARTILE(range, 1) function)
- third quartile Q_3 (using the = QUARTILE(range, 3) function)
- maximum time (using the = MAX(range) function)

and then the inter-quartile range (IQR):

$$IQR = Q_3 - Q_1$$

Finally the whisker limit values were calculated, such that a values below the lower limit or above the upper limit could be identified as possible outliers (see introduction):

$$w_{\min} = Q_1 - 1.5 \times IQR$$

$$w_{\max} = Q_3 + 1.5 \times IQR$$

Tip: you can swap rows and columns in Excel by copying and doing "paste special" with the Transpose option.

			min	wmin	Q1	median	Q3	wmax	max
2000	Men	swim	1064	1036	1090	1098	1126	1180	1164
		cycle	3459	3473	3532	3552	3571	3630	3812
		run	1854	1780	1928	1995	2027	2176	2219
		overall	6504	6380	6571	6648	6699	6891	6972
	Women	swim	1143	1082	1185	1229	1254	1358	1365

		cycle	3929	3667	3941	4020	4124	4398	4482
		run	2063	2020	2226	2275	2364	2571	2612
		overall	7241	6945	7399	7524	7701	8155	8201
2004	Men	swim	1069	1063	1087	1097	1103	1127	1174
		cycle	3644	3361	3703	3771	3931	4273	4177
		run	1912	1767	1978	2029	2119	2331	2375
		overall	6668	6355	6800	6944	7097	7542	7534
	Women	swim	1117	1048	1161	1187	1236	1349	1320
		cycle	4138	3784	4194	4247	4467	4877	4812
		run	2053	1967	2228	2306	2402	2662	2626
		overall	7483	7181	7652	7757	7966	8437	8559
2008	Men	swim	1080	1070	1091	1101	1106	1128	1136
		cycle	3468	3504	3529	3536	3546	3571	3559
		run	1846	1696	1939	2004	2101	2343	2319
		overall	6533	6376	6631	6693	6801	7057	7011
	Women	swim	1189	1149	1195	1202	1226	1273	1263
		cycle	3834	3698	3857	3924	3963	4122	4064
		run	1997	1882	2138	2206	2309	2566	2521
		overall	7108	6883	7292	7407	7564	7973	7819

A visual inspection of this table shows 13 values that would be classed as outliers using the standard definition as per the above formula.

The two that are "too small" are both in cycling, as discussed above, and the difference in time is quite small:

Men's cycling 2000, $\frac{3552}{3459} = 1.027$, the winner was 2.7% faster than the median

Men's cycling 2004, $\frac{3536}{3468} = 1.020$, the winner was 2% faster than the median

The difference in speed does not seem impossible: it is not sufficient to suggest that these points should be discarded. Similar values are known to occur, e.g. 1969 Tour de France, Stage 17: Eddy Merckx was 1.9% faster than the median.

The values that are "too large" probably imply that an athlete was completely exhausted, even possibly injured, and failed to keep up with the others. There is no reason to believe that these values are "wrong" in any sense and it seems reasonable to keep and use them in the analysis.

Looking at the actual values in Excel, one could remove these values from the 2000 and 2004 data sets if one used only the first 36 competitors (overall ranking) in year and gender. (Of course, this might make the IQR smaller and other values might then be called into question).

Looking at the 2008 data however, the "outlier" man who took 1136 seconds in the swimming came 14th overall and it would clearly be ridiculous to discard his data (he was just 3.2% slower than the swimming median speed).

I will therefore keep all these values rather than discarding them.

The data is tabulated in Appendix 1.

(b) Plotting the data in Autograph

Columns of data were selected in Excel and then pasted into the "Enter raw data" (1D statistics pages) or "Enter XY data" (2D graph pages) data entry boxes. Prior to the paste, the new dataset was given a name e.g. "2000 Men".

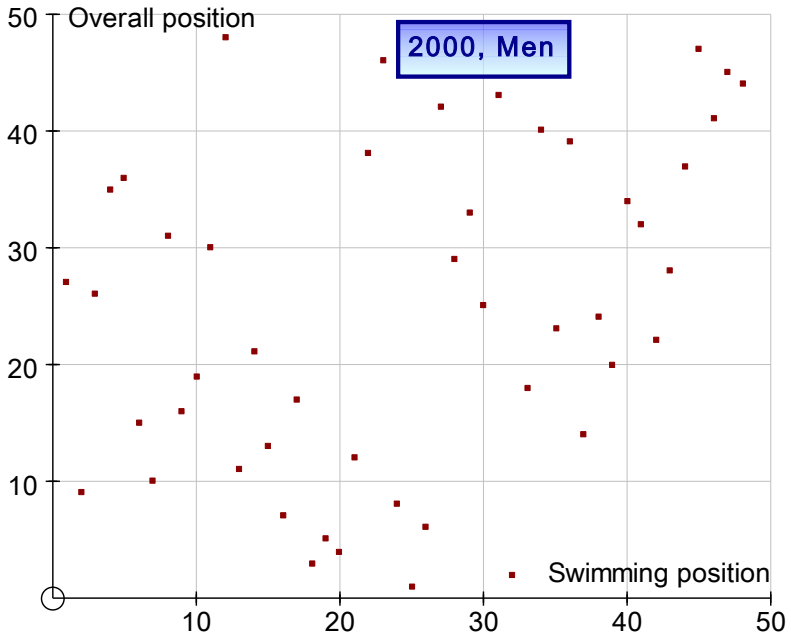
The axis font size was set at 14 to stop it being too small when the graph was later pasted into Word. The page was resized to make it more nearly square: this is easier to read and interpret.

Each time a box and whisker diagram was created, the data name was visible in the status bar at the bottom of the window. A text box was immediately added beside the left-hand whisker so that the box could be identified and to avoid confusion with other boxes. As discussed in the introduction, a subset of the data was used to avoid bias in the median values:

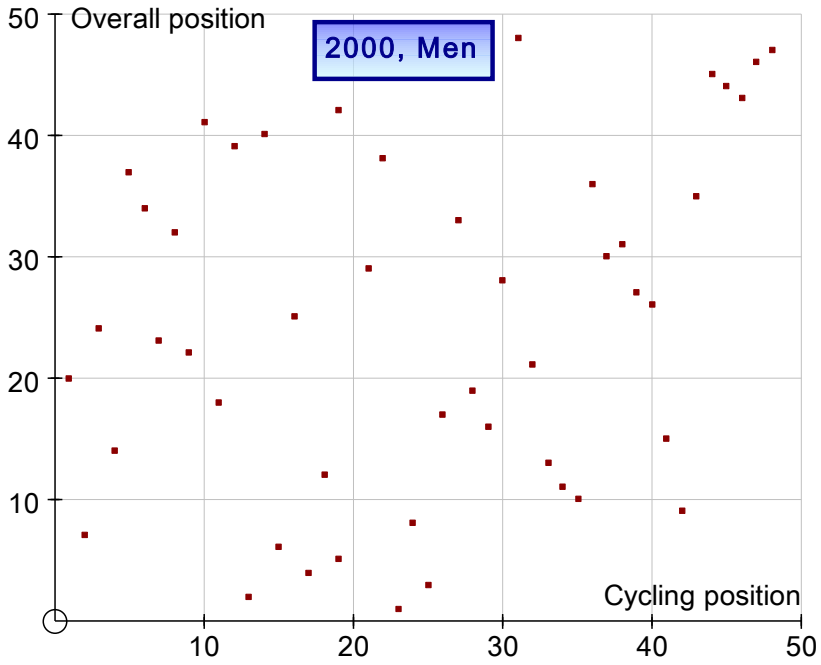
Year/MW	2000-M	2000-W	2004-M	2004-W	2008-M	2008-W
Total athletes	48	40	45	44	50	45
Use the overall first:	40	40	44	44	45	45

First hypothesis diagrams

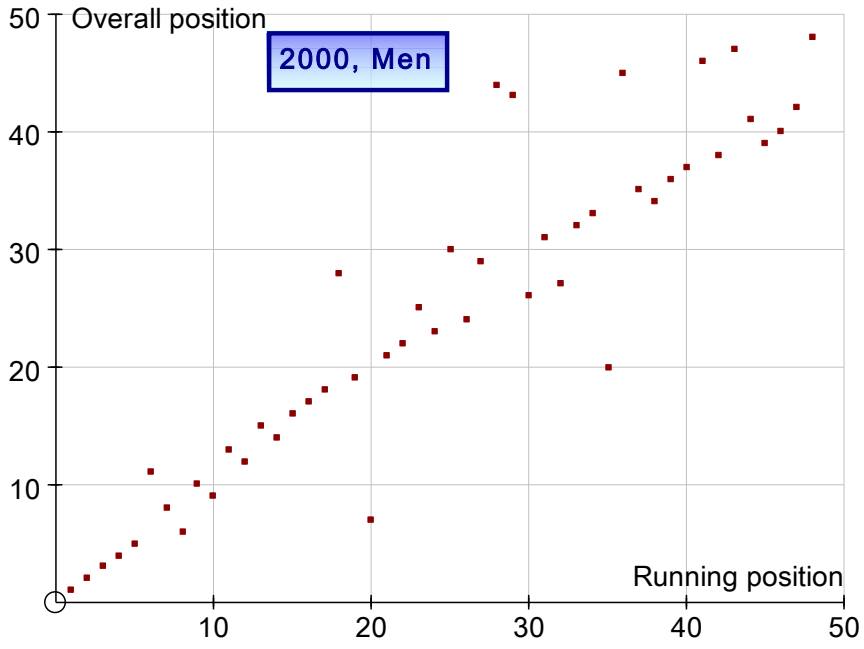
All graphs and diagrams are for the men in year 2000 unless otherwise stated.



Spearman's ranking coefficient = 0.34, a very weak correlation (using "show statistics" in Autograph).

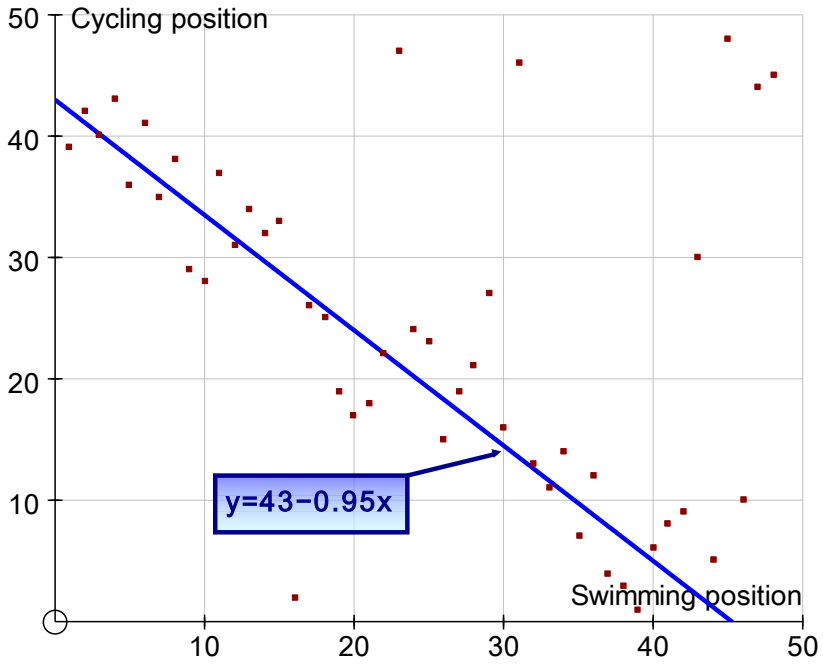


Spearman's ranking coefficient = 0.27, a very weak correlation.

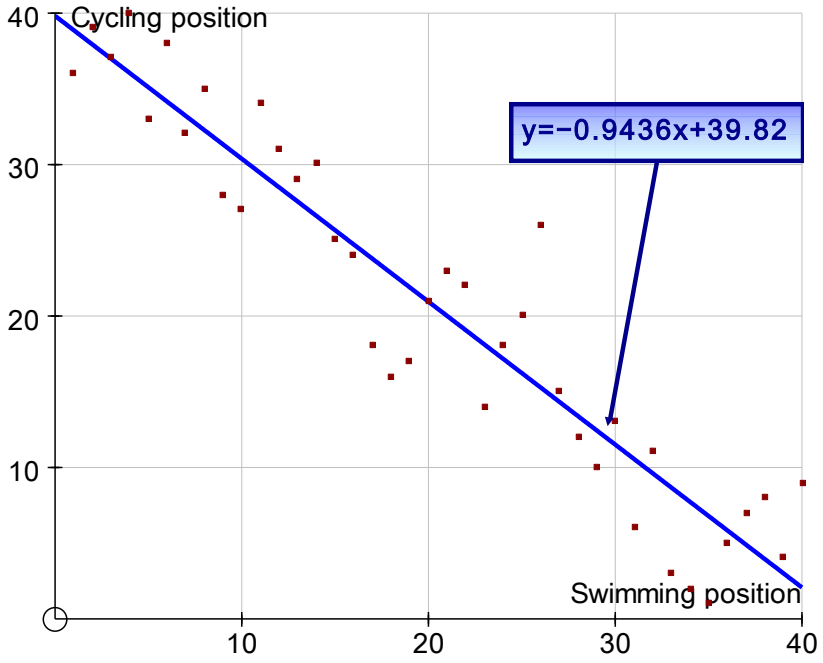


Spearman's ranking coefficient = 0.93, a very strong correlation.

First hypothesis, second investigation



Spearman's ranking coefficient = -0.46, a negative correlation. The best fit line has been added by eye to ignore the few points that are far from the main trend. Since this line equation is for 48 data points it will be difficult to make a prediction for a year with fewer competitors so (as discussed in the introduction) I will replot it after regenerating the rankings in Excel for just the best 42 (overall) , minus Olivier Marceaux and Oscar Galindez, who are then the two points that do not follow the general trend, hence 40 data points. The line here is a first order (i.e. straight) best fit regression line generated automatically in Autograph:

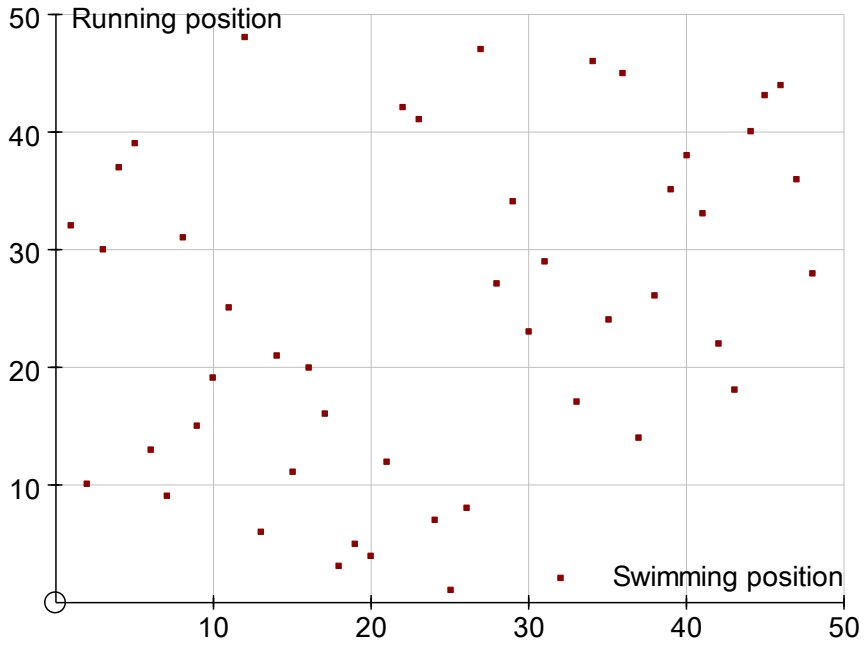


Clearly there is some scatter around the line but the mean of several points should show less scatter. I predict that for the other data sets, taking the best 40 (overall) in each set:

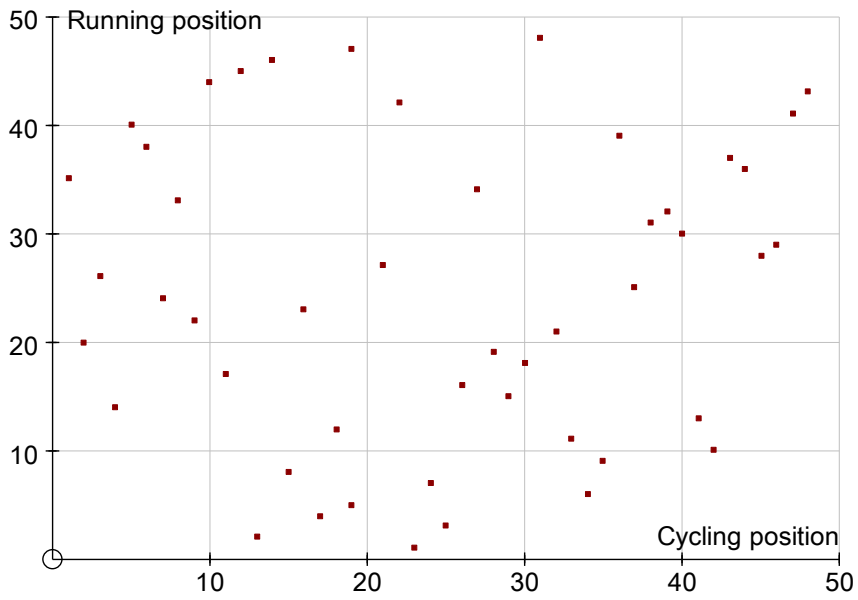
- the best 5 swimmers (mean position $\frac{1+2+3+4+5}{5}=3$) will have a mean cycling position of $-0.9436\times 3+39.82=36.9892$ (37 rounded to 2 significant figures)
- the worst 5 swimmers (mean position $\frac{36+37+38+39+40}{5}=38$) will have a mean cycling position of $-0.9436\times 38+39.82=3.9632$ (4.0 rounded to 2 significant figures)

These will be calculated and compared later.

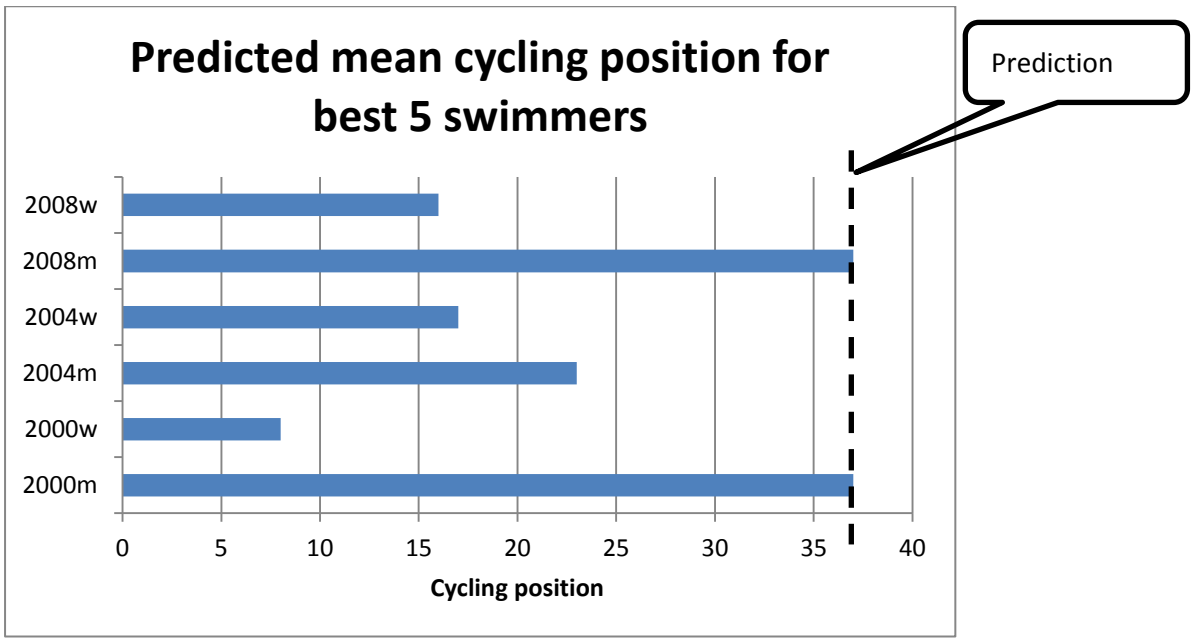
Returning to the full year 2000 men data set:

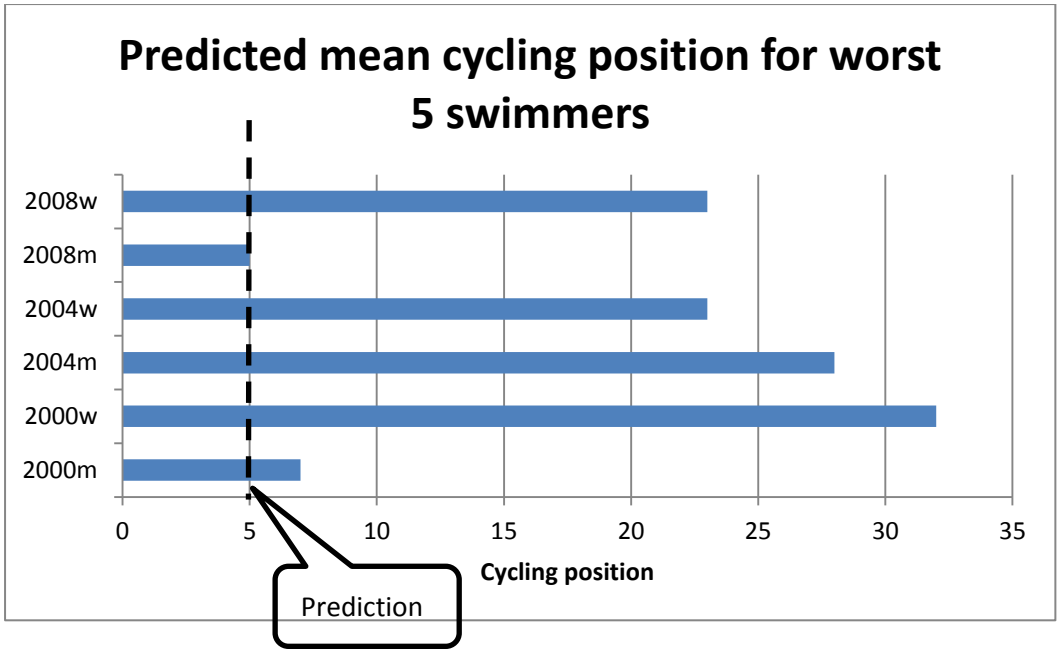


Spearman's ranking coefficient = 0.28, a very weak correlation.



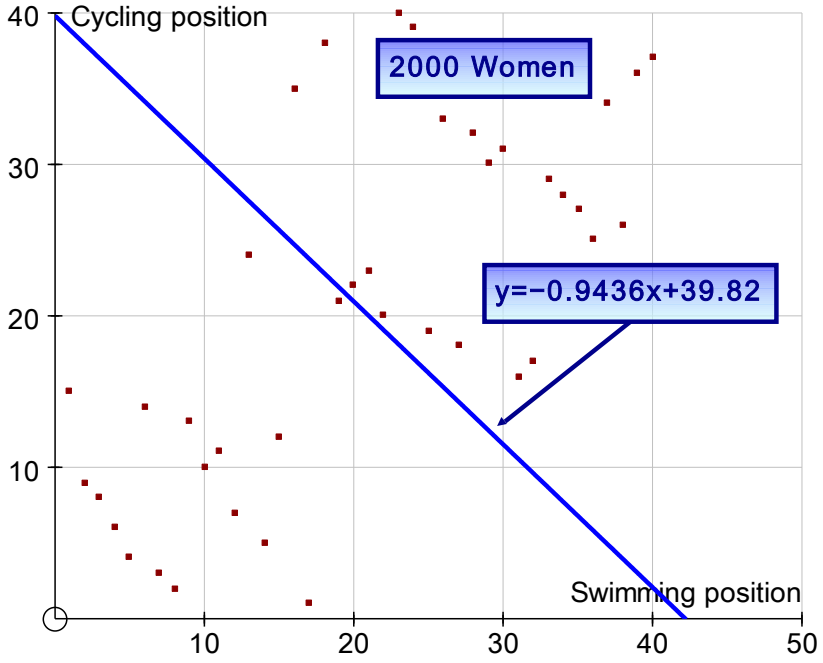
Spearman's ranking coefficient = 0.04, no c orrelation.



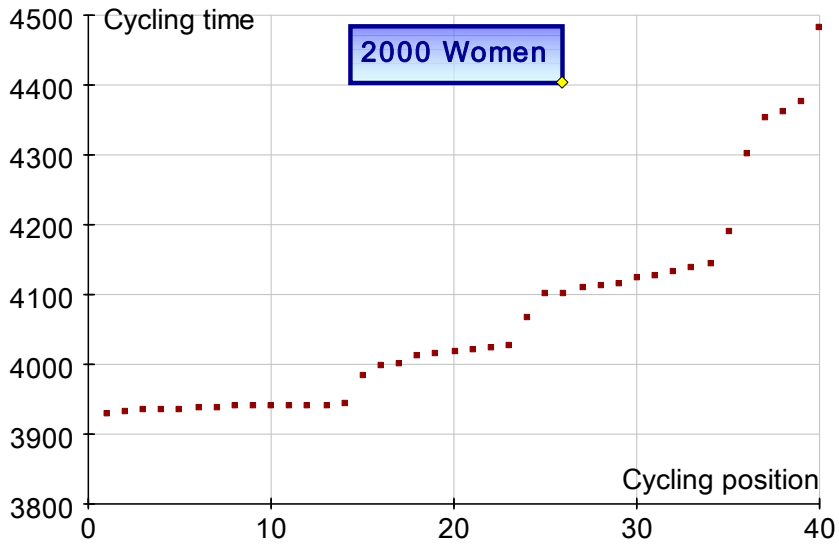


Clearly the best fit line formula does not succeed in predicting the mean cycling position for the best 5 and worst 5 swimmers in each competition, though it is close for the men in 2000 and 2008. (It is not exact even in 2000 because the line fit is for the whole data not just the first and last 5 points).

Plotting a scatter graph for the worst-predicted case above (Women in 2000) shows that there is a great deal of scatter with "waves" of points that each seem to follow some trend:

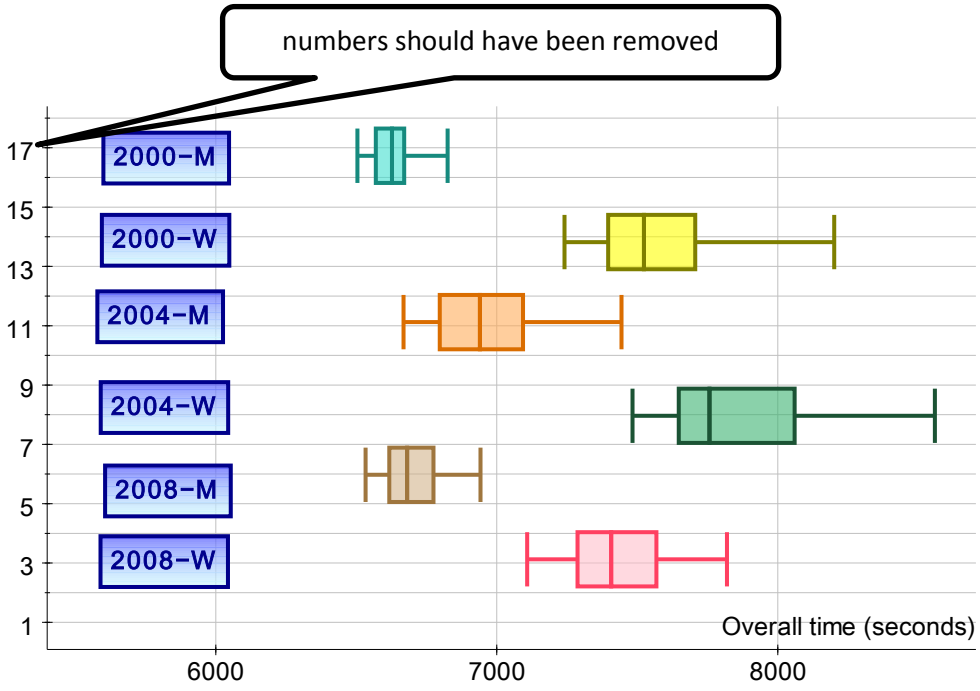


The best-fit line equation used in the prediction above would predict the position of those around 20th place but not the best or the worst competitors. Plotting the actual finish times of each woman in this cycling event shows that they were travelling in groups that arrived within a short time period: the position within this group is perhaps a poor indicator of overall ability:

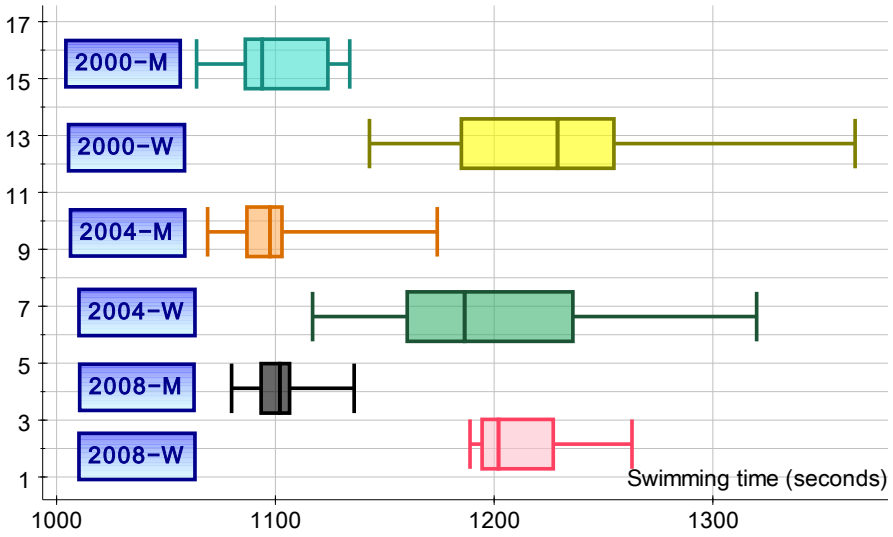


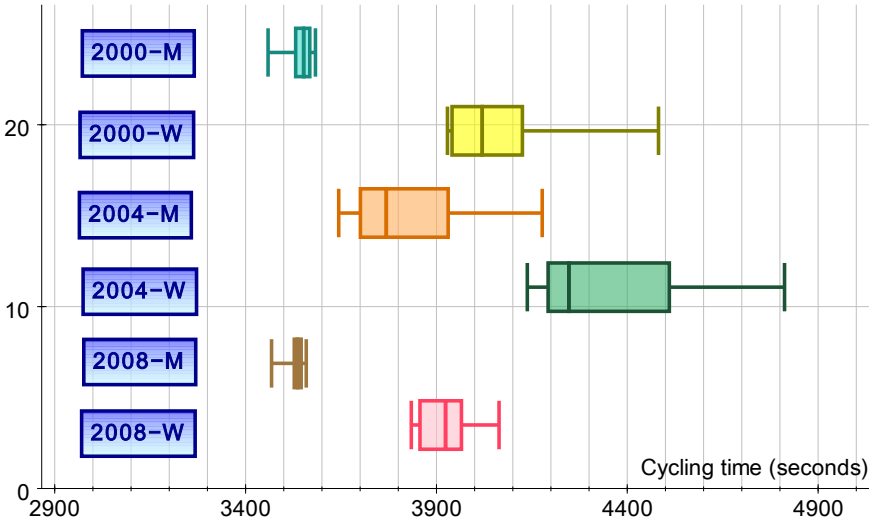
Second hypothesis diagrams.

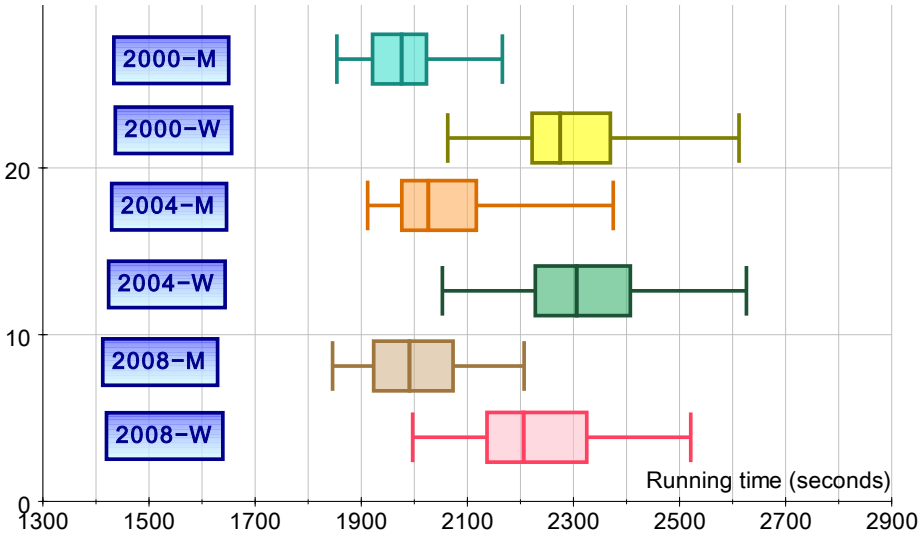
Main investigation – overall times



Further investigation – individual events:


2000-M
2000-W
2004-M
2004-W
2008-M
2008-W
Swimming time (seconds)
1000
1100
1200
1300


2000-M
2000-W
2004-M
2004-W
2008-M
2008-W
Cycling time (seconds)
2900
3400
3900
4400
4900


2000-M
2000-W
2004-M
2004-W
2008-M
2008-W
Running time (seconds)
1300
1500
1700
1900
2100
2300
2500
2700
2900

Median times (seconds)

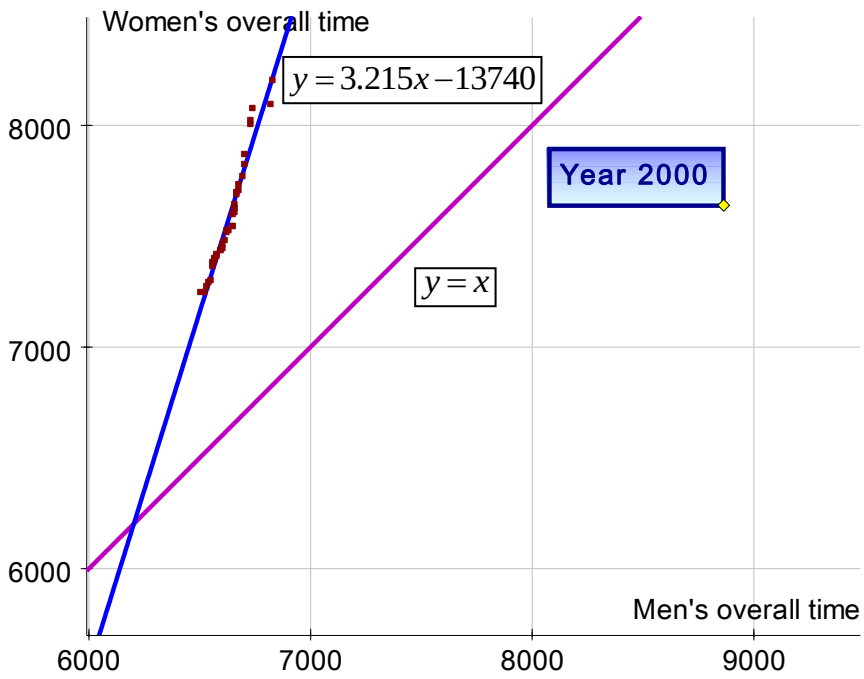
	Overall	Swimming	Cycling	Running
2000 Men	6627	1094	3552	1976
2000 Women	7524	1229	4020	2275
2004 Men	6940	1098	3769	2027
2004 Women	7757	1187	4247	2306
2008 Men	6681	1102	3536	1991
2008 Women	7407	1202	3924	2206

ratio of median times, women/men:	Overall	swimming	cycling	running
2000	1.14	1.12	1.13	1.15
2004	1.12	1.08	1.13	1.14
2008	1.11	1.09	1.11	1.11
Mean	1.121	1.098	1.123	1.132

Inter-quartile range of times (units: seconds).

	Overall	Swimming	Cycling	Running
2000 Men	101	38	37	102
2000 Women	310	70	184	148
2004 Men	296	16	230	141
2004 Women	413	76	317	179
2008 Men	157	13	18	150
2008 Women	281	33	109	188

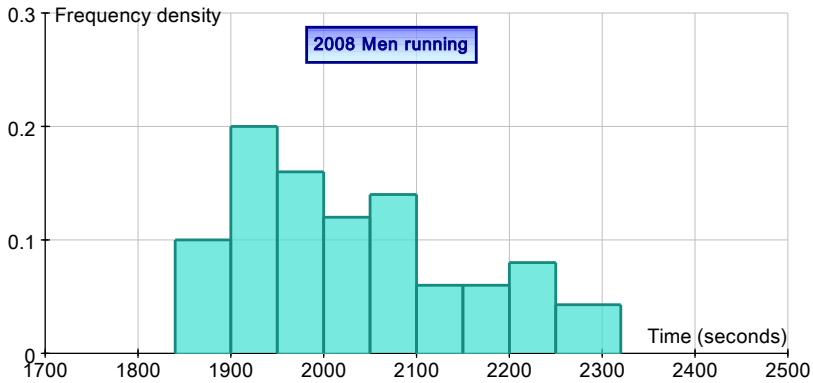
Since the overall interquartile range times for women were significantly larger than for men, I decided to pursue a further investigation as to the possible causes. To do this I plotted a scatter graph in which each point used the overall time data for (1^{st} man, 1^{st} woman), (2^{nd} man, 2^{nd} woman) etc:



The gradient of about 3 for the best fit line on this graph is unexpected. A possible conjecture is that there are few women triathletes, whereas male triathletes are drawn from a much bigger pool. Looking at the first (for instance) 15 competitors to arrive after the leader, they arrived at a rate of 4.4 athletes/minute (women) but 9.2 athletes/minute (men) – more than twice as rapidly. Overall, the men's arrival rate was over three times faster than the women's. One might imagine that there are about 3 times as many male athletes in each ability band as there are female athletes. Conversely, there may be some other explanation altogether.

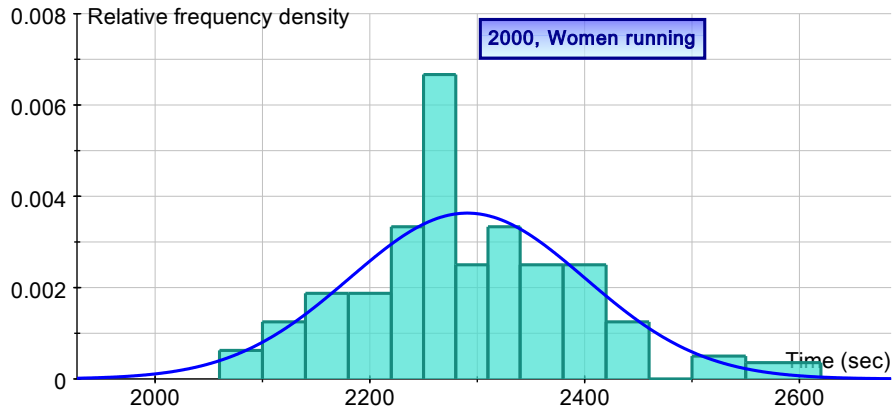
Third hypothesis

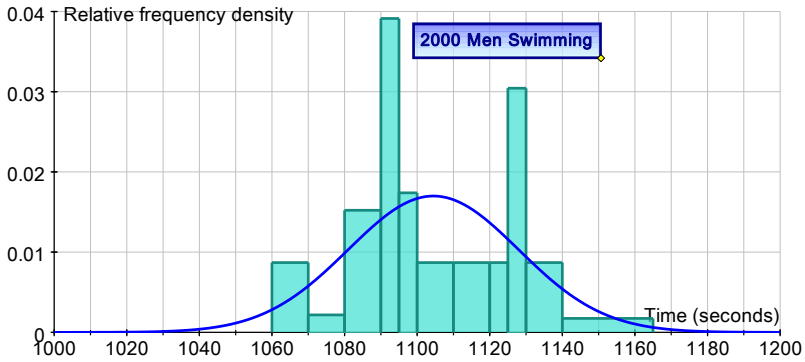
Many of the histograms do show a positive skew (tail to right), for instance:

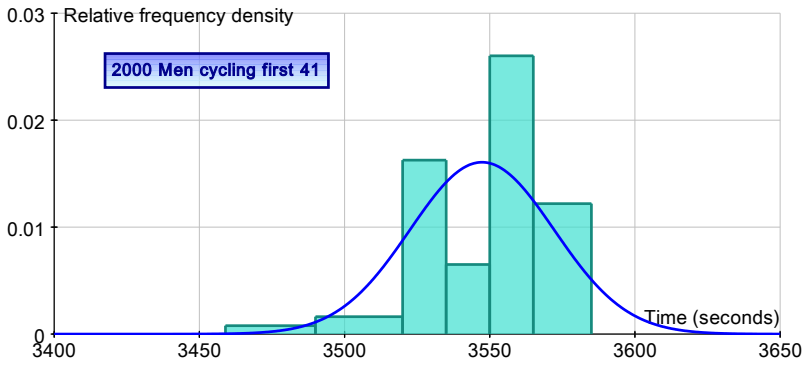


This supports the hypothesis.

I have chosen a quota sample of three events (a running, a swimming and a cycling) to compare against the Normal distribution:





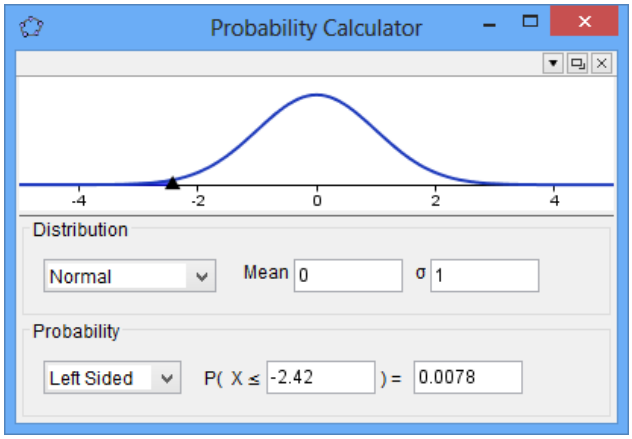


Looking at these three histograms (out of 18 possibilities for individual events), only the first is even vaguely like the Normal curve. The peaks in the swimming and cycling histograms show where many contestants finished with very similar times – an observer watching the race would have noticed an exciting photo-finish as they arrived together.

In the 2000 Men's cycling, 5 men took far longer than the others, giving a separate peak off to the right of this graph (around 3700 seconds). To avoid such people affecting the calculation of the mean and standard deviation in Excel, I will leave out any people finishing much later in each race than the main body (plotting a scatter graph of each column in Excel to identify such people). This data is in a new Excel workbook "Best 40 in each, scrambled" indicating that each column has been sorted independently.

		minimum	Mean Time	Standard deviation	mean-2sd	(mean-min)/sd	Number < (m-2sd)
2000 M	S	1063.89	1102.007	20.58511	1060.837	1.85	0
	C	3458.6	3547.32	24.81803	3497.684	3.57	2
	R	1853.73	1979.262	60.26916	1858.724	2.08	1
2004 M	S	1069	1095.442	10.94022	1073.561	2.42	1
	C	3644	3798.976	111.3823	3576.212	1.39	0
	R	1912	2028.225	71.88758	1884.45	1.62	0
2008 M	S	1080	1096.524	8.033377	1080.457	2.06	1
	C	3468	3531.72	20.47832	3490.763	3.11	3
	R	1846	2030.32	125.375	1779.57	1.47	0
2000 W	S	1142.78	1219.967	36.391	1147.185	2.12	1
	C	3928.5	4014.924	76.86031	3861.204	1.12	0
	R	2063.03	2262.82	82.54021	2097.74	2.42	1
2004 W	S	1117	1188	41.99942	1104.001	1.69	0
	C	4138	4326.372	172.9416	3980.489	1.09	0
	R	2053	2319.023	128.4825	2062.058	2.07	1
2008 W	S	1189	1201.714	11.22134	1179.272	1.13	0
	C	3834	3913.591	60.05767	3793.476	1.33	0
	R	1997	2236.222	125.0612	1986.1	1.91	0

Looking at the right-hand column in this table, we see that actually 11 people finished this far before the mean time for their race. I can conclude that the Normal distribution is not a good model for predicting probability in Triathlon competitions. Even if I had set the limit at 2.42 standard deviations (as 2004 Men swimming in table) I would have still had 5 people before the limit – and for a Normal distribution, this corresponds to a probability of only 0.0078 per event (Geogebra applet):



Conclusions

First hypothesis

The scatter graphs of overall position versus position in each sport disprove the first hypothesis.

- swimming and overall position: poor correlation (r = 0.34)
- cycling and overall position: poor correlation (r=0.27)
- running and overall position: very strong correlation (r=0.93)

I had expected similar Spearman's rank correlation coefficients for each event. The fact that running has a very strong correlation coefficient (0.93) shows that to do well overall you must do well in the running; the other two sports are much less important to the overall outcome.

The negative correlation between swimming and cycling is perhaps to be expected if one sport uses arms and the other leg muscles. This means that, with hindsight, the first hypothesis is inevitably false because you will not get athletes who are equally good at all disciplines: those who are good at swimming will always tend to be poor at cycling.

I have used the trend line on the 40-athlete swimming versus cycling graph to predict the cycling outcome for the other years, both men and women, and have compared this by looking at a bar graph of the mean positions in cycling for the best and worst 5 swimmers. The agreement between data and prediction is quite poor for most events. This can be explained to some extent by the tendency for cyclists to group together.

Second hypothesis

The first part of this, that women are uniformly slower than men, seems to be supported. One can see this in the box and whisker plots, comparing for each event the men's times (in 2000, 2004, 2008) and the equivalent women's times

The table showing the ratio women's time/ men's time indicates that women are typically 10% slower at swimming, 12% slower at cycling and 13% slower at running.

Comparing the inter-quartile ranges however shows that there in much more scatter in the women's times than expected. This part of the hypothesis is therefore disproved.

The scatter graph of women's time against men's time for pairs of competitors (1^{st}, 1^{st}) etc shows that (considered in terms of overall time) the men arrive at the finishing line at approximately 3× the rate (per minute) of the women. This might suggest that many more men train as triathletes and so the best ones are more similar in terms of ability than the equivalent women.

Third hypothesis

Statistics for 2000 Women running

Raw Data Statistics:	
Number in sample, n:	40
Mean, x:	2290.72
Standard Deviation, x:	109.844
Range, x:	549
Lower Quartile:	2222
Median:	2275
Upper Quartile:	2369.5
Semi I.Q. Range:	73.75

Grouped Data Statistics:	
Total Frequency, n:	40
Mean, x:	2291.88
Standard Deviation, x:	106.546
Modal Class:	2250-
Lower Quartile:	2227.5
Median:	2276.25
Upper Quartile:	2360
Semi I.Q. Range: 66.25	

Statistics for 2008 W running

Raw Data Statistics:	
Number in sample, n:	45
Mean, x:	2236.22
Standard Deviation, x:	125.061
Range, x:	524
Lower Quartile:	2137
Median:	2206
Upper Quartile:	2325
Semi I.Q. Range: 94	

Appendix

Mens 2000 Olympic Triathlon data.

Place	Triathlete	Swimming	Cycling	Running	Total time*
1	Simon Whitfield (CAN)	00:18:18	00:59:12	00:30:54	01:48:24
2	Stephan Vuckovic (GER)	00:18:36	00:58:52	00:31:10	01:48:38
3	Jan Řehula (CZE)	00:18:12	00:59:14	00:31:21	01:48:47
4	Dmitriy Gaag (KAZ)	00:18:13	00:59:08	00:31:42	01:49:04
5	Ivan Rana (ESP)	00:18:12	00:59:10	00:31:48	01:49:11
6	Miles Stewart (AUS)	00:18:20	00:59:00	00:31:54	01:49:15
7	Olivier Marceau (FRA)	00:18:12	00:58:12	00:32:54	01:49:18
8	Reto Hug (SUI)	00:18:17	00:59:12	00:31:52	01:49:21
9	Simon Lessing (GBR)	00:17:45	00:59:40	00:32:00	01:49:24
10	Timothy Don (GBR)	00:18:01	00:59:31	00:31:57	01:49:29
11	Andriy Glushchenko (UKR)	00:18:10	00:59:30	00:31:50	01:49:30
12	Andreas Raelert (GER)	00:18:13	00:59:10	00:32:08	01:49:31
13	Martin Krňávek (CZE)	00:18:11	00:59:22	00:32:05	01:49:38
14	Leandro Macedo (BRA)	00:18:46	00:58:40	00:32:24	01:49:51
15	Volodymyr Polikarpenko (UKR)	00:18:00	00:59:37	00:32:15	01:49:52
16	Craig Watson (NZL)	00:18:05	00:59:19	00:32:38	01:50:02
17	Hunter Kemper (USA)	00:18:12	00:59:15	00:32:39	01:50:06
18	Markus Keller (SUI)	00:18:38	00:58:51	00:32:46	01:50:15
19	Carl Blasco (FRA)	00:18:06	00:59:18	00:32:54	01:50:18
20	Conrad Stoltz (RSA)	00:18:47	00:57:39	00:33:58	01:50:24
21	Takumi Obara (JPN)	00:18:10	00:59:21	00:32:58	01:50:30
22	Juraci Moreira (BRA)	00:18:49	00:58:45	00:33:10	01:50:45
23	Eneko Llanos (ESP00:)	00:18:44	00:58:44	00:33:20	01:50:48

Swimming	Cycling	Running	Total time*
1098	3552	1854	6504
1116	3532	1870	6518
1092	3554	1881	6527
1093	3548	1902	6544
1092	3550	1908	6551
1100	3540	1914	6555
1092	3492	1974	6558
1097	3552	1912	6561
1065	3580	1920	6564
1081	3571	1917	6569
1090	3570	1910	6570
1093	3550	1928	6571
1091	3562	1925	6578
1126	3520	1944	6591
1080	3577	1935	6592
1085	3559	1958	6602
1092	3555	1959	6606
1118	3531	1966	6615
1086	3558	1974	6618
1127	3459	2038	6624
1090	3561	1978	6630
1129	3525	1990	6645
1124	3524	2000	6648

Swimming order	Cycling order	Running order	Overall order
25	23	1	1
32	13	2	2
18	25	3	3
20	17	4	4
19	19	5	5
26	15	8	6
16	2	20	7
24	24	7	8
2	42	10	9
7	35	9	10
13	34	6	11
21	18	12	12
15	33	11	13
37	4	14	14
6	41	13	15
9	29	15	16
17	26	16	17
33	11	17	18
10	28	19	19
39	1	35	20
14	32	21	21
42	9	22	22
35	7	24	23

24	Jean-Christophe Guinchard (SUI)	00:18:47	00:58:37	00:33:27	01:50:51	1127	3517	2007	6651	38	3	26	24
25	Ryan Bolton (USA)	00:18:31	00:59:07	00:33:15	01:50:53	1111	3547	1995	6653	30	16	23	25
26	Hamish Carter (NZL)	00:17:48	00:59:37	00:33:32	01:50:57	1068	3577	2012	6657	3	40	30	26
27	Craig Walton (AUS)	00:17:44	00:59:37	00:33:37	01:50:58	1064	3577	2017	6658	1	39	32	27
28	Oscar Galindez (ARG)	00:18:51	00:59:21	00:32:48	01:50:59	1131	3561	1968	6659	43	30	18	28
29	Johannes Enzenhofer (AUT)	00:18:23	00:59:12	00:33:28	01:51:02	1103	3552	2008	6662	28	21	27	29
30	Csaba Kuttor (HUN)	00:18:07	00:59:34	00:33:25	01:51:06	1087	3574	2005	6666	11	37	25	30
31	Stephan Bignet (FRA)	00:18:02	00:59:36	00:33:34	01:51:12	1082	3576	2014	6672	8	38	31	31
32	Alessandro Bottoni (ITA)	00:18:49	00:58:44	00:33:45	01:51:18	1129	3524	2025	6678	41	8	33	32
33	Vassilis Krommidas (GRE)	00:18:26	00:59:16	00:33:47	01:51:29	1106	3556	2027	6689	29	27	34	33
34	Peter Robertson (AUS)	00:18:47	00:58:41	00:34:11	01:51:39	1127	3521	2051	6699	40	6	38	34
35	Joachim Willen (SWE)	00:17:49	00:59:43	00:34:09	01:51:41	1069	3583	2049	6701	4	43	37	35
36	Hideo Fukui (JPN)	00:17:57	00:59:32	00:34:35	01:52:05	1077	3573	2075	6725	5	36	39	36
37	Gilberto Gonzalez (VEN)	00:18:54	00:58:40	00:34:38	01:52:13	1134	3520	2078	6733	44	5	40	37
38	Ben Bright (NZL)	00:18:15	00:59:12	00:34:50	01:52:17	1095	3552	2090	6737	22	22	42	38
39	Armando Barcellos (BRA)	00:18:46	00:58:52	00:36:05	01:53:43	1126	3532	2165	6823	36	12	45	39
40	Nick Radkewich (USA)	00:18:44	00:58:55	00:36:06	01:53:45	1124	3535	2166	6825	34	14	46	40
41	Matias Brain (CHI)	00:18:57	00:58:46	00:36:02	01:53:45	1137	3526	2162	6825	46	10	44	41
42	Eric van der Linden (NED)	00:18:22	00:59:10	00:36:59	01:54:32	1102	3550	2219	6872	27	19	47	42
43	Rob Barel (NED)	00:18:35	01:03:32	00:33:30	01:55:37	1115	3812	2010	6937	31	46	29	43
44	Jan Knobelauch Hansen (DEN)	00:19:24	01:02:49	00:33:29	01:55:42	1164	3769	2009	6942	48	45	28	44
45	Roland Melis (AHO)	00:19:21	01:02:48	00:34:03	01:56:12	1161	3768	2043	6972	47	44	36	45
46	Hiroyuki Nishiuchi (JPN)	00:18:16	01:03:57	00:34:47	01:57:00	1096	3837	2087	7020	23	47	41	46
47	Mikhail Kuznetsov (KAZ)	00:18:55	01:04:39	00:35:39	01:59:13	1135	3879	2139	7154	45	48	43	47
48	Dennis Looze (NED)	00:18:08	00:59:21	00:42:55	02:00:24	1088	3561	2575	7224	12	31	48	48