MAT022 Foundations of Statistics and Data Science

Summative Assessment 2019/20

Charles Wills

c1977808

02/01/2020

Contents

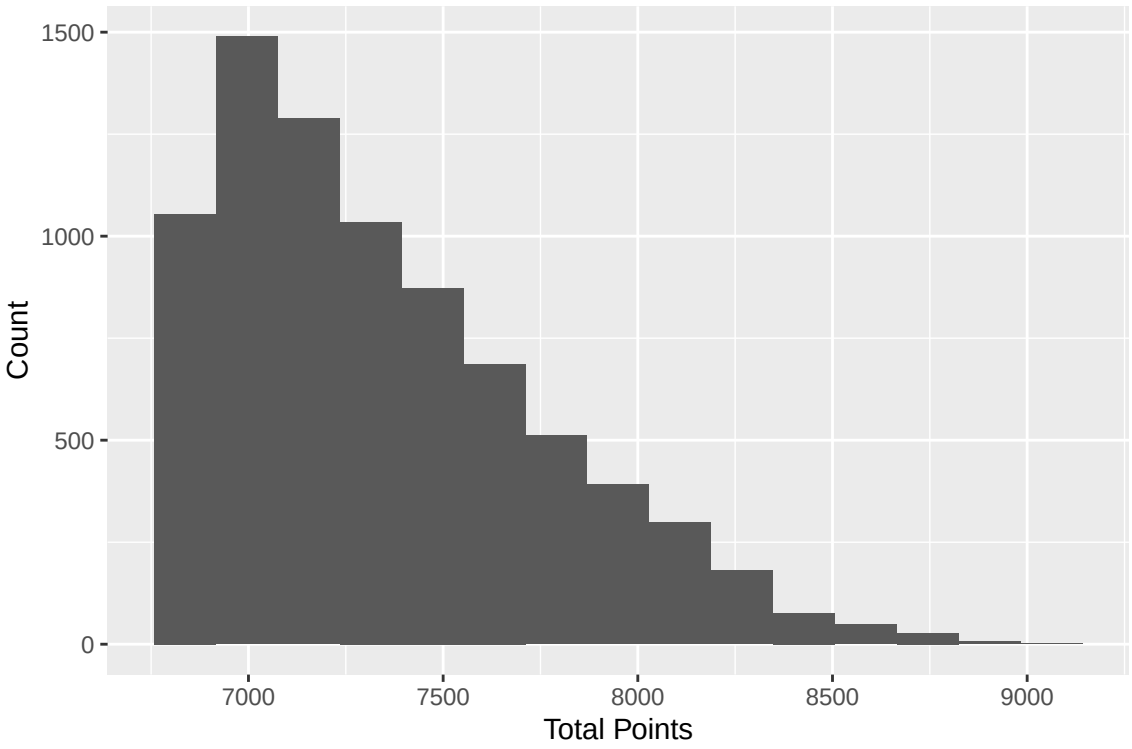
Introduction

"You, sir, are the greatest athlete in the world"

King Gustav of Sweden declared when he handed the winner's prize to Jim Thorpe at the 1912 olympic games in Stockholm. A bold statement maybe, but can you argue with his logic? Thorpe had proven himself to be the best athlete across ten different sports. It's harder to be great at ten things than one, therefore he must be the best athlete in the world. In this paper we will investigate if decathletes can lay stake to this claim, and investigate which continent produces the best decathletes.

The decathlon dataset records the performance of decathletes between 1986 and 2006. It has 7,968 observations with data for 2,709 unique decathletes. The points and result for each of the ten events is given and the total points column indicates their performance across all events. The person with the highest total points after the ten events is crowned the winner. The average total points tally, by median, is 7255 with half of all athlete's scoring between 7020 and 7349. The histogram of total points shows a positive skew, where we see more observations on the right tail of the graph. These are the athlete's with the highest points tally, who are likely to have won the decathlon.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6800    7020    7255    7349    7608    9026
```



Decathletes vs Olympic Medalists

Jack of all trades or master of none? To answer this we need to compare the decathletes performance with professionals in respective events. If a decathlete can claim to be the best athlete in the world, they should be at least as good as athletes that focus only on one event. Should the data show there is no statistical significance between results of decathletes and professionals, then it's fair to conclude King Gustav was correct.

To make this comparison we'll need an additional dataset. We'll use an olympic track and field dataset from Kaggle. This has the result of every athletic event at the olympic games between 1896 and 2016. This data was originally web scraped from the official olympic website. It's important to briefly discuss potential limitations with this data:

- It was compiled by a third party. I have reviewed the code used to collate the data, but it's still possible errors have occurred during the process.
- It includes results at olympic games, these only occur every four years. This leads to a smaller sample than the decathlon dataset offers when comparing similar time periods.
- It covers a different range of years than the decathlon dataset.
- It only records the results of the medalists (gold, silver and bronze). This means the results are for only the three best performers in each event.

To overcome these limitations we'll need to make a few adjustments:

- Only analyse years that are in both the decathlon and olympic datasets i.e. 1985 - 2006. Note: Due to the olympic data containing fewer observations we'll actually use the year range of 1984 - 2008 to ensure our sample size is sufficient.
- Extract the top three performers from each year in the decathlon dataset. This ensures we're comparing elite performers in both datasets. Otherwise we'd be comparing only olympic medalists with all decathletes, which may bias the results.

We'll start with the 100m sprint, testing if decathletes are significantly slower than professional sprinters. To do this we'll use a two-sample t-test. This compares the mean of two groups and tests to see if they are significantly different from each other. The table below shows the average 100m time for the decathletes and the professionals. It's clear the professionals yield a better result, they average a time that's over two seconds faster than the decathletes, but is this difference statistically significant?

athlete_type	mean	sd	n
decathlete	12.010294	0.0975364	68
professional	9.945556	0.1275203	18

Step 1 - Define the hypothesis

It's important we do this before consulting the data. If not, we risk potentially biasing the results by choosing parameters that fit our agenda. Here we are trying to determine if decathletes are significantly slower than professional sprinters in the 100m. Therefore we will test the following hypothesis:

$H_0 : \mu_D = \mu_P$ (The mean 100m time is the same for the decathletes as it is for the professional sprinters.)

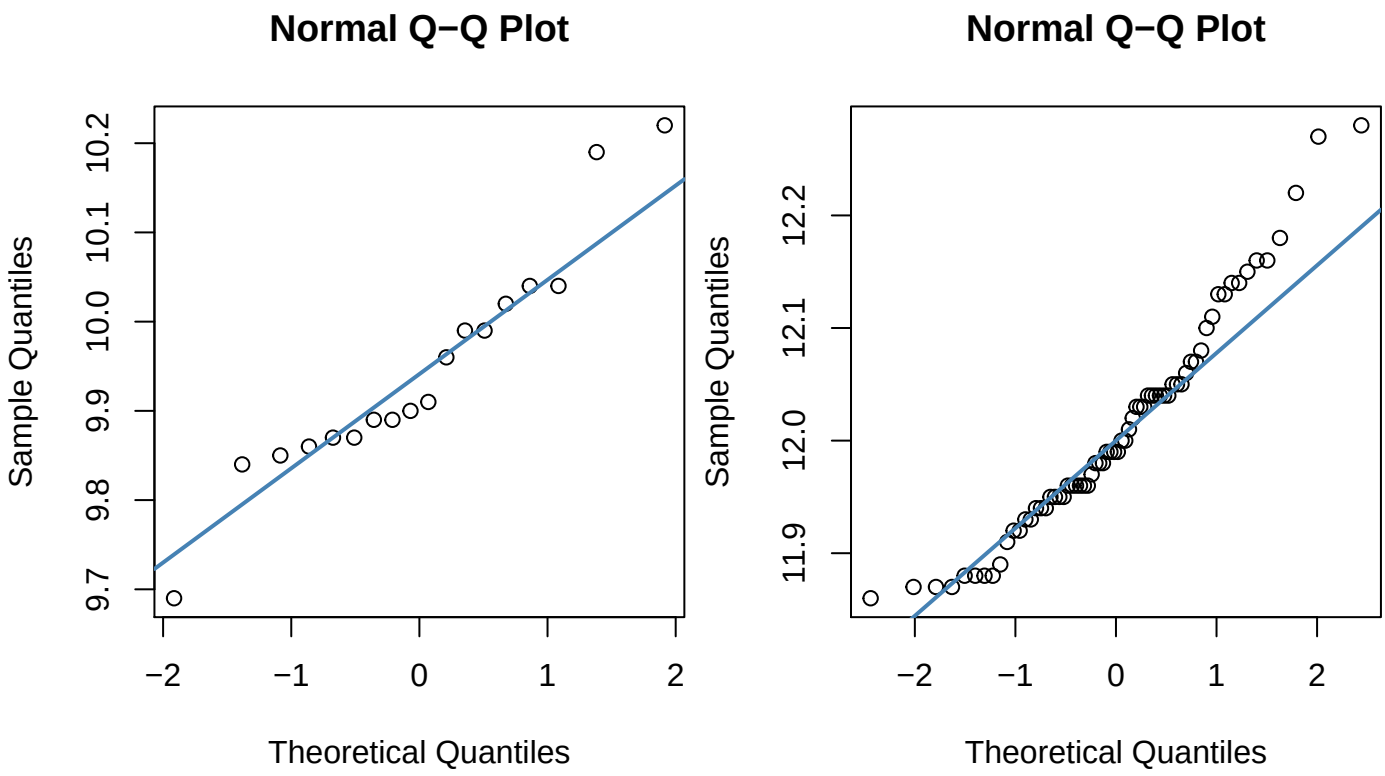
$H_1 : \mu_D > \mu_P$ (The mean 100m time is greater for the decathletes than it is for the professional sprinters.)

Step 2 - Construct the test

We're going to perform a one-tailed, two-sample t-test using the conventional 5% significance level. The t-test compares the means from two independent samples and assumes that:

- The measurements in each population follow a normal distribution.
- The populations the samples came from have equal variance.

Let's check these two assumptions hold true for our data. To check for normality, a simple Q-Q (Quantile-Quantile) plot can be used.



The points on both plots lie close to the blue line, which represents a theoretical normal Q-Q plot. Therefore we can be satisfied the first assumption of normality holds true. To check the populations of both samples have equal variance we can perform Levene's test. This is also a hypothesis test, therefore we need to define the hypothesis and then run the test:

$H_0 : \sigma_D^2 = \sigma_P^2$ (Variances **are** equal between samples)

$H_1 : \sigma_D^2 \neq \sigma_P^2$ (Variances **are not** equal between samples)

Using $\alpha = 0.05$ and the classical Levene's procedure whereby the distance from the mean, rather than median, is used in the calculation. As discussed by Brown and Forsythe (1974, pp. 364-367), the mean is a good choice when you have symmetric and moderately tailed distributions and when the underlying shape of the distribution is known.

```
##
##  Classical Levene's test based on the absolute deviations from the mean
##  ( none not applied because the location is not set to median )
##
## data:  combined_df$result
## Test Statistic = 1.558, p-value = 0.2154
```

The resulting p-value of the test is 0.215 (3dp). As this is greater than our significance level of 0.05 we do not have sufficient evidence to reject H_0. We can therefore assume the variances are equal between the samples which means the second assumption is satisfied.

Step 3 - Perform the test

```
##
##  Two Sample t-test
##
## data:  result by athlete_type
```

```
## t = 74.681, df = 84, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.018756      Inf
## sample estimates:
##   mean in group decathlete mean in group professional
##                  12.010294                   9.945556
```

Step 4 - Conclusion

The resulting p-value is far smaller than our significance level (0.05). This means we can strongly reject H_0 in favour of H_1 and conclude that the decathletes are significantly slower at running the 100m than the professional olympic sprinters. The 95% confidence interval suggests the true difference in 100m time between decathletes and professional sprinters is more than two seconds. This probably isn't a surprising result, winning an olympic medal in the 100m sprint is no mean feat.

So can we conclude that King Gustav was wrong when he declared Jim Thorpe the "greatest athlete in the world"? Well, sprinting is just one area of the decathlon, so let's perform a similar test for two completely different decathlon events; high jump (jumping) and discus (throwing).

The high jump and discus satisfy the same assumptions as above and yield the following mean, standard deviation, and sample size.

Discus	Mean	Std Dev	Sample Size
decathlete	51.149	1.393	66
professional	67.131	1.684	18

High Jump	Mean	Std Dev	Sample Size
decathlete	2.170	0.031	83
professional	2.346	0.017	17

Using the same hypothesis and significance level as above, we get the following results:

Event	p-value (3dp)	Lower Conf Int	Upper Conf Int
Discus	0.000	-Inf	-15.336
High Jump	0.000	-Inf	-0.163

It turns out the decathletes average significantly lower scores than the olympic athletes. For the high jump, the true difference between the decathletes and Olympians mean score is more than 16cm. For the discus, the true difference is more than 15 metres. Therefore it's probably fair to say King Gustav was incorrect, or at least not entirely correct. While Jim Thorpe was a fantastic athlete, he wouldn't have stood much of a chance when running, jumping, or throwing against olympic champions. Sorry, Jim.

Decathletes by Continent

We've seen that decathletes can't compete with Olympians in the individual events that make up the decathlon. But how do they compare amongst themselves? More specifically, who produces the best decathletes? To answer this, we could compare multiple countries with one another, but with 107 different countries it could

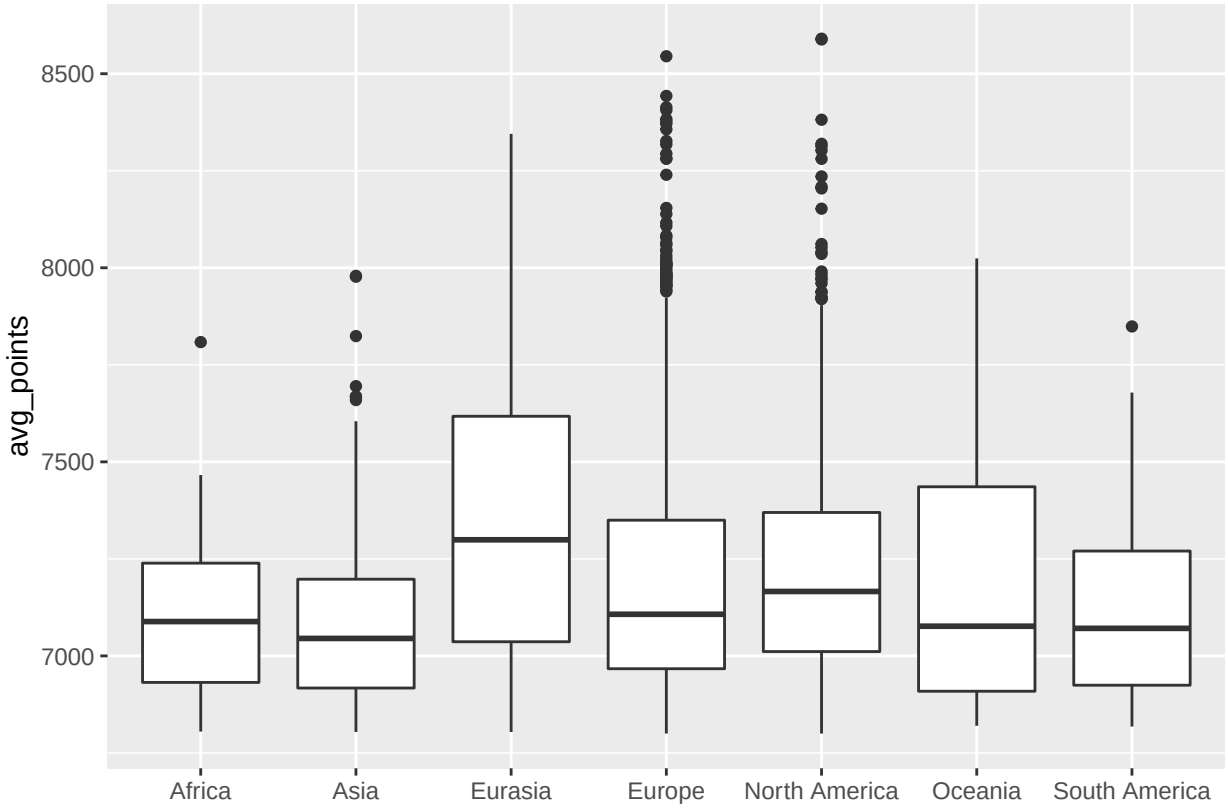
take a while. A better way to answer this question is by grouping each country by continent. The question then becomes; which continent produces the best decathletes?

To answer this we need a country to continent lookup table, this has been sourced from the internet. Similarly to the Olympic data this has been compiled by a third party which needs to be considered with respect to accuracy. We can then use this lookup to get the continent each decathlete represents. A new set of problems arise;

- Decathletes have represented countries that are no longer officially recognised. For example; the Soviet Union and East and West Germany. In an attempt not to lose data these countries have been mapped to the most appropriate continent based on their new country name. For example West Germany maps to Germany which maps to Europe.
- Countries that span multiple continents, namely Europe and Asia, meant it impossible to classify as one or the other. All these cases occurred in Eurasia which has subsequently formed it's own group, despite not being an officially recognised continent.

As we're looking to compare the means between three or more groups, that suggests a one-way ANOVA test may be sensible. The assumptions are similar to those of the t-test we performed earlier. Firstly, the underlying distributions are assumed to be normal. As seen in Introduction the Total Points variable is not normally distributed, it's positively skewed. Therefore we aren't able to use one-way ANOVA and will instead need to use a non-parametric alternative which makes no assumption about the underlying distribution.

The Kruskal-Wallis rank sum test is considered the non-parametric equivalent of one-way ANOVA. It assumes each group (ie continent) has the same shape. This can be checked with a simple boxplot which confirms the assumption holds.



It also requires independence of observations within and between groups. That is, no decathlete can appear more than once in the data. As it stands we are violating this condition, many decathletes have competed over multiple years which means the observations are not considered independent. To overcome this we can average each decathletes score across all their decathlons and that points total will form an observation.

Therefore we now have 2709 observations, one for each unique decathlete. As with all statistical tests we will define the hypothesis and construct the test before consulting the data.

Step 1 - Define the hypothesis

The Kruskal-Wallis test assigns a rank based on the observation and calculates the median rank for each group. Therefore we test whether the median of all groups are equal:

$H_0 : \eta_1 = \eta_2 = ... = \eta_7$ (The median rank of all continents are equal.)

$H_1 : \eta_i \neq \eta_j$ for some $i \neq j$ (At least two continents have a different median rank.)

Step 2 - Construct the test

We will use the conventional significance level of $\alpha = 0.05$. The Kruskal-Wallis test statistic can be written as:

$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(N+1)$

where;

- N is the total number of observations.
- T_i is the rank sum for the ith group.
- n_i is the median of group i.

It can be shown that if the null hypothesis is true then H is approximately distributed like a random variable from a Chi-square distribution on k-1 degrees of freedom. Therefore the critical value can be calculated by $\chi_6^2 = 12.592$ (with $\alpha = 0.05$ and rounded to 3dp). If our test statistic exceeds this value we can reject the null hypothesis.

Step 3 - Perform the test

```
##
##  Kruskal-Wallis rank sum test
##
## data:  avg_points by continent
## Kruskal-Wallis chi-squared = 83.154, df = 6, p-value = 7.959e-16
```

Step 4 - Conclusion

Our test statistic (83.154) greatly exceeds our critical value (12.592) which means we can strongly reject H_0 in favour of H_1. This is also confirmed by our p-value being far smaller than our significance level of 0.05. We can interpret this as at least two continents show a significant difference in points scored. This means that at least one continent is better than one other continent. But we can't conclude yet which continents these are. For that we have to perform post hoc tests.

Step 5 - Post hoc tests

For context the average points by continent table below shows how each continent compares. This doesn't tell us if the difference in points scored are statistically significantly different from one another. To do this we need to perform pairwise comparisons.

```
## # A tibble: 7 x 4
##   continent      mean    sd     n
##   <fct>         <dbl> <dbl> <int>
## 1 Asia          7084.  217.   238
## 2 Africa        7108.  220.    39
## 3 South America 7110.  228.    43
## 4 Oceania       7188.  324.    62
## 5 Europe        7194.  312.  1422
## 6 North America 7229.  311.   683
## 7 Eurasia       7354.  386.   222
```

To compare continents we could run several pairwise comparisons using the Mann-Whitney sum of ranks test. This is the non-parametric equivalent of the two sample t-test. However, this can prove time consuming when comparing seven groups. Therefore, a more suitable post hoc test is Connover & Iman's procedure for multiple comparisons.

Another consideration we have to make is around unintentionally inflating the risk of Type I error. If we are making multiple comparisons between groups using significance level $\alpha = 0.05$ we'd only have to make 20 comparisons before we would expect to incorrectly reject H_0. Therefore we have to make an adjustment to our significance level. For the Conover & Iman procedure we will use the Bonferonni correction. This is calculated by dividing the original significance level by the number of comparisons we are making. We have seven continents which means we'll be making 21 comparisons. Therefore our corrected significance level is $0.05 \div 21 = 0.002$ (3dp).

To be explicit our hypotheses for these comparisons are:

$H_0 : \eta_i = \eta_j$

$H_1 : \eta_i \neq \eta_j$

Conover and Iman's multiple comparisons yields the following p-values for each combination of continents.

```
##
##  Pairwise comparisons using Conover's-test for multiple
##                           comparisons of independent samples
##
## data:  avg_points by continent
##
##               Africa Asia    Eurasia Europe North America Oceania
## Asia          1.0000 -       -       -      -             -
## Eurasia       0.0033 3.6e-15 -       -      -             -
## Europe        1.0000 2.2e-05 5.6e-08 -      -             -
## North America 0.4830 1.7e-09 0.0055  0.0305 -             -
## Oceania       1.0000 1.0000  0.0123  1.0000 1.0000        -
## South America 1.0000 1.0000  0.0015  1.0000 0.3328        1.0000
##
## P value adjustment method: bonferroni
```

This can be interpreted such that a p-value between two continents that is less than our Bonferroni corrected p-value (0.002) suggests a significant difference between the continents performance. From the table above these are; Asia and Eurasia, Asia and Europe, Asia and North America, Eurasia and Europe, Eurasia and South America.

It can be difficult to summarise multiple comparisons tests succinctly but this seems to suggest that Asia perform the worst at decathlon. They record significantly lower points totals than three other continents. While Asia have performed well in recent years in track and field events it could be argued this was a result of Beijing hosting the 2012 summer Olympics. Before this time the investment in Asian track and field events was more limited, which could explain this finding.

Meanwhile, Eurasia seem to be the continent that produces the best decathletes, showing statistically significant gaps to Asia, Europe, and South America. Towards the end of the cold war the Soviet Union, which makes up a high proportion of the Eurasia group, had a lot of success in track and field events and plenty of money was invested in it. This may explain why they appear to be the most successful continent in the decathlon.

reference this article https://statistics.laerd.com/spss-tutorials/kruskal-wallis-h-test-using-spss-statistics.php