

# Some Truly Olympic Activities



Mary Mortlock    [marym@harker.org](mailto:marym@harker.org)  
The Harker School

Matt Carlton    [mcarlton@calpoly.edu](mailto:mcarlton@calpoly.edu)  
Cal Poly State University

<http://statweb.calpoly.edu/mcarlton/olympics>

# Some Truly Olympic Activities

## CONTENTS

### **Olympic Swimming**

*data collection from the web, descriptive statistics, comparing distributions*

### **Olympic Men's Marathon**

*regression...lots of regression...*

### **The Statistics Olympics!**

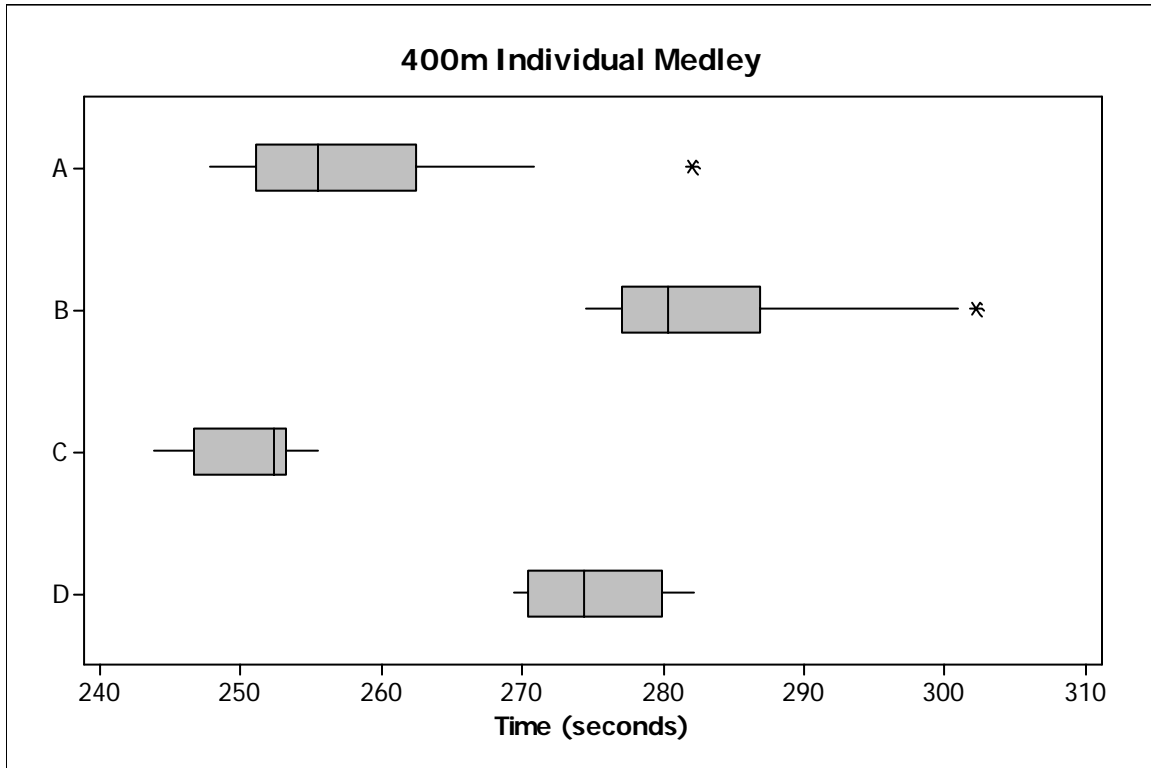
*in-class data collection, hypothesis testing galore*

### **Bonus Material**

*wait and see...*

# Olympic Swimming

Consider the following graph.



These graphs show the distributions of the swimmers' times for the men's and women's 400m individual medley at the 2008 Olympics. Two show the results of all the heats, and two show the finals (with just 8 swimmers in each). Identify which of the four graphs (A, B, C, D) is...

- I. Women's heats
- II. Women's final
- III. Men's heats
- IV. Men's final

Explain your reasoning!

If you were shown just one of these graphs, would you be able to tell whether there were more than 8 swimmers in the race? Why or why not?

Go to the following website:

<http://www.2008.nbcolympics.com/swimming/resultsandschedules/>

This lists the times of all the swimming events, both heats and finals, from the 2008 Olympics held in Beijing, China.

You can categorize an event in 3 ways.

Stroke:           **Breaststroke, Backstroke, Butterfly, Freestyle, Individual  
Medley or Relay**  
Gender:           **Male or Female**  
Type of Race:   **Heat or Final**

Pick two events to compare by changing just one of these three features. For instance, if you chose the 100m men's freestyle final, you could choose the 100m women's freestyle final (to compare gender), or the 100m men's freestyle heats (to compare type of race) - make sure you download them all, or the 100m men's butterfly final (to compare stroke).

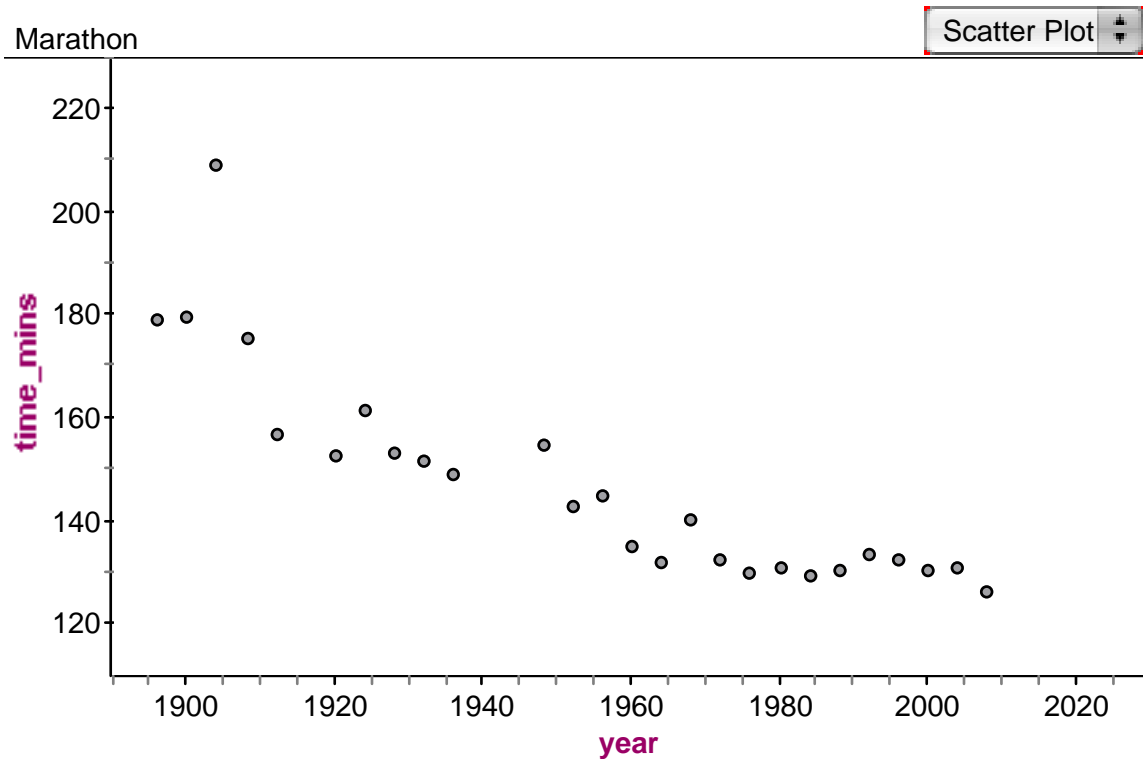
Convert these times into seconds, and put the times into List 1 and List 2 on your TI.

Create parallel box plots of the times, and write a few sentences comparing and contrasting the two distributions. Remember to consider shape, center, and spread, and use the context of the situation.

# Olympic Men's Marathon

Open the data file that contains the winning times of the Men's Marathon from all the Summer Olympic Games that have been held since 1896, the first year of the modern Olympics.

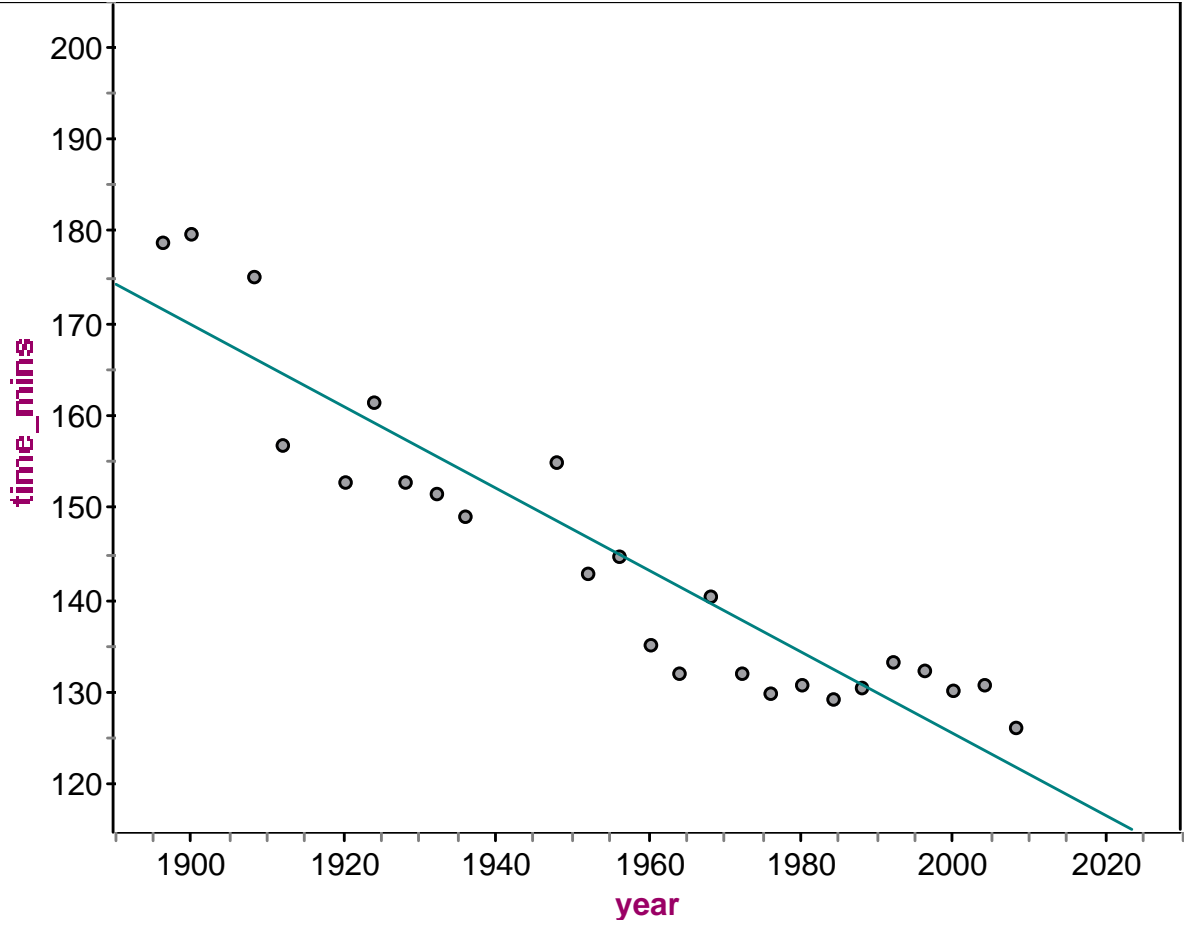
- (a) Before creating a scatter plot, decide which of the variables year and time would be the response variable and which would be the predictor variable.
- (b) Make a scatter plot showing a line of least squares.
- (c) Look at the plot. Is the association between time and year positive or negative?
- (d) What happened in 1904?
- (e) Run a complete regression analysis with the data for 1904 removed. What is the regression equation? Use variable names, not  $x$  and  $y$ .
- (f) Predict the winning marathon time when the year is 0. What constant is this? Does this make sense in the context of the question?
- (g) Using the least squares line, find the *expected* winning times in 1896 and in 2000. Divide the change in time by the change in year. This gives the average increase/decrease in time for every extra year. What constant is this? Does this make sense in the context of the question?
- (h) The Olympics weren't held in 1916, 1940, and 1944. Use the regression equation to predict the winning times in those years if the Olympics had been held.
- (i) What will be the winning time in 2288? Why do some of these times [the ones in (h)] make more sense than others [the one you just calculated in part (i) and the one in part (f)]?
- (j) Calculate the residuals for the years 1964 and 1968. How many non-negative residuals are there between the years 1960 and 1984 (7 Olympics in total)?



American Thomas Hicks was leading in the second half of the 1904 St. Louis marathon when, exhausted from the 90 degree weather and single water stop, he asked to lie down; instead his handlers repeatedly fed him strychnine and egg whites and then helped the hallucinating Hicks across the finish line in a time of 3:28:53.

Marathon - 1904

Scatter Plot



$$\hat{y} = 1007.5 - 0.4411x \quad (r^2 = 0.87)$$

# Some Cautions about Correlation and Regression

Correlation and regression are useful tools, but they should not be used blindly!!! If the relationship between the variables is not linear, then these methods are not very helpful. As always, plot the data first!

Other cautions:

## 1. Extrapolation

Use of the regression line to make predictions outside the range of observed data is called extrapolation. Usually, this is not a good idea. Though a linear relation may be true for the observed range of values, this may not be true outside this range. Remember the shepherds in the Nike sandals, running the marathon in year 0?

## 2. Unusual Observations

If a few observations don't fit the general pattern, think about why this might be. Can you discover a reason for this unusual behavior? Why was the winning marathon time so high in 1904?

## 3. Other Variables

The relationship between two variables may be influenced by other variables. Another variable could have an important effect on the relationship among the variables in a study.

## 4. Association is Not Causation

Strong association is not enough to make statements about causation!

## Data for Olympic Men's Marathon

Year	Time (mins)
1896	178.8
1900	179.7
1904	208.9
1908	175.3
1912	156.9
1920	152.9
1924	161.4
1928	153.0
1932	151.6
1936	149.3
1948	154.9
1952	143.0
1956	145.0
1960	135.3
1964	132.2
1968	140.4
1972	132.3
1976	129.9
1980	131.0
1984	129.3
1988	130.5
1992	133.4
1996	132.6
2000	130.2
2004	130.9
2008	126.2

# The Statistics Olympics!

Congratulations on being selected to participate in the Statistics Olympics! You will proudly represent your nation in the Balance and Flexibility events. Gold, Silver and Bronze medals will be awarded to the winners of each event. Prepare to win!

The Lithuanian team and the British team will compete against each other in two events. Additionally, you will be using the data from these events to answer several research questions.

**REQUIREMENTS:** Each team needs a tape measure, a stop watch, and a table for recording data.

---

## **EVENT #1: BALANCE**

Instructions for each team member: Stand on one leg with your eyes closed. You must have one leg off the ground and keep the other one still. Have a teammate record the time (in seconds) from the moment you start until you hop, touch or grab onto something, or fall over!

---

## **EVENT #2: FLEXIBILITY**

Instructions for each team member: Sit with your back against a wall and your legs flat on the floor in front of you. With your toes at 90° to the floor, try and stretch your fingers as far past your toes as possible. Have a teammate measure the distance you can stretch past your toes, in centimeters. If you reach exactly to your toes, a distance of 0 should be recorded. A negative value signifies that you failed to reach your toes.

After you have been measured once, take a few seconds to stretch out. Then follow the same instructions and get measured again. Both readings should be taken before moving on to the next athlete.

---

## Statistics Olympics: Analysis

Answer each question using the appropriate hypothesis test. Reminder: Explain your results so that the journalists covering this event from all over the world can interpret the results for their viewers.

1. In the Statistics Olympics, we compare nations by their average score on an event. Are the results for the Flexibility event (using the *first tries only*) significantly different for your nation and the other nation?
2. Calculate the proportion of Lithuanians that have a positive Flexibility 1 score and the proportion of Brits that have a positive Flexibility 1 score. [Note: 0 is not positive!] Test the hypothesis that the proportion of Lithuanians that can reach further than their toes is different from the corresponding proportion of Brits.
3. An Olympics "expert" believes that healthy, sober people can balance on one leg with eyes closed for 15 seconds, on average. Some of the members of the International Olympic Committee believe that 15 seconds just doesn't sound right. Use the Balance data from both nations' athletes to determine whether the mean differs significantly from the belief of the expert. What report would you make to the International Olympic Committee?
4. Rather than look to see if the mean is different than 15 seconds, the IOC decide to look at the proportion of young healthy students that can balance over 15 seconds. Use both nations' Balance data to find if more than half of people can balance over 15 seconds. Has your report changed?
5. Is there a relationship between flexibility and balance? Make a scatter plot using the Balance and Flexibility 1 data from all the athletes, and test to see if the relationship is significant.
6. Do athletes improve their Flexibility with a second try? Using the Flexibility 1 and Flexibility 2 data from all the athletes, test to see if the flexibility improves, on average, between the first and second try.
7. Consider the test used for #2. See if you can answer the same question using another type of hypothesis test. Do you come up with the same conclusion? Is one test better (more appropriate) than the other? Why?
8. Come up with another question you can answer using the Statistics Olympics data, one not already addressed in #1-#7 above. Conduct the proper analysis.

## Statistics Olympics: Mary & Matt's Classroom Data

Great Britain			Lithuania		
<u>Balance</u>	<u>Flex 1</u>	<u>Flex 2</u>	<u>Balance</u>	<u>Flex 1</u>	<u>Flex 2</u>
40	-1	2	45	0	3
24	0	3	43	0	2
25	1	5	31	0	4
4	2	4	46	4	6
37	3	7	28	7	9
43	4	7	10	8	11
5	6	7	19	0	3
2	7	10	14	-3	4
37	7	12	19	13	17
33	7	11	7	11	16
4	8	12	23	-6	-2
2	-6	-3	17	6	14
3	-6	-3	45	2	0
7	-2	1	49	2	8
4	-2	2	8	1	5
16	11	14	11	2	2
19	12	15	12	6	11
10	-6	-3	14	5	8
23	-6	3	24	6	9
17	-2	1	23	6	10
18	2	0	35	12	15
5	6	9	32	9	14
11	7	10	20	18	24
9	11	13	17	20	22
13	10	15			

## Statistics Olympics: Correct hypotheses

1. Let  $\mu_1$  = the average Flexibility score for all Lithuanians and let  $\mu_2$  = the average Flexibility score for all Britons.

$$H_0: \mu_1 = \mu_2$$

$$H_a: \mu_1 \neq \mu_2$$

2. Let  $p_1$  = the proportion of all Lithuanians who can stretch beyond their toes and let  $p_2$  = the proportion of all Britons who can stretch beyond their toes.

$$H_0: p_1 = p_2$$

$$H_a: p_1 \neq p_2$$

3. Let  $\mu$  = the true average time someone can balance on one leg.

$$H_0: \mu = 15$$

$$H_a: \mu \neq 15$$

4. Let  $p$  = the proportion of all people who can balance on one leg for over 15 seconds.

$$H_0: p = 0.5$$

$$H_a: p > 0.5$$

5. Let  $\beta$  = the true slope relating Flexibility and Balance scores for the population.

$$H_0: \beta = 0$$

$$H_a: \beta \neq 0$$

6. Let  $\mu_d$  = the population average difference in Flexibility scores, with differences computed as Flexibility 2 - Flexibility 1.

$$H_0: \mu_d = 0$$

$$H_a: \mu_d > 0$$

7. A chi-square test (of homogeneity) will give the same result. In particular, the chi-square statistic for the 2x2 table will be the square of the z-statistic from #2, and the P-values will be identical.

8. One example is to compare mean Balance scores between Lithuania and Great Britain.

# Data Analysis for Mary & Matt's Data

(see "Statistics Olympics: Analysis" for the original questions)

Note: This page only contains the "number-crunching" portion of the analysis, as provided by Minitab. Students should always identify parameters, write hypotheses, check necessary conditions, crunch the numbers, then give a conclusion.

---

#1.

## Two-Sample T-Test and CI: Flex 1, Nation

Two-sample T for Flex 1

Nation	N	Mean	StDev	SE Mean
GB	25	2.92	5.76	1.2
LI	24	5.38	6.26	1.3

Difference =  $\mu$  (GB) -  $\mu$  (LI)  
Estimate for difference: -2.45500  
95% CI for difference: (-5.91799, 1.00799)  
T-Test of difference = 0 (vs not =): T-Value = -1.43 P-Value = 0.160 DF = 46

---

#2.

## Test and CI for Two Proportions: Flex1Pos, Nation

Nation	X	N	Sample p
GB	16	25	0.640000
LI	18	24	0.750000

Difference =  $p$  (GB) -  $p$  (LI)  
Estimate for difference: -0.11  
95% CI for difference: (-0.365762, 0.145762)  
Test for difference = 0 (vs not = 0): Z = -0.84 P-Value = 0.404

---

#3.

## One-Sample T: Balance

Test of  $\mu = 15$  vs not = 15

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Balance	49	20.4694	13.6582	1.9512	(16.5463, 24.3925)	2.80	0.007

---

#4.

## Test and CI for One Proportion: Over15s

Test of  $p = 0.5$  vs  $p > 0.5$

Variable	X	N	Sample p	95% Lower Bound	Z-Value	P-Value
Over15s	29	49	0.591837	0.476346	1.29	0.099

#5.

### Regression Analysis: Flex 1 versus Balance

The regression equation is  
Flex 1 = 4.22 - 0.0050 Balance

Predictor	Coef	SE Coef	T	P
Constant	4.225	1.591	2.65	0.011
Balance	-0.00501	0.06487	-0.08	0.939

S = 6.13854    R-Sq = 0.0%    R-Sq(adj) = 0.0%

*Note: For this analysis, either variable can be the predictor.*

#6.

### Paired T-Test and CI: Flex 2, Flex 1

Paired T for Flex 2 - Flex 1

	N	Mean	StDev	SE Mean
Flex 2	49	7.53061	6.36757	0.90965
Flex 1	49	4.12245	6.07465	0.86781
Difference	49	3.40816	1.96764	0.28109

95% lower bound for mean difference: 2.93671

T-Test of mean difference = 0 (vs > 0): T-Value = 12.12    P-Value = 0.000

#7.

### Chi-Square Test: Pos, NotPos

Expected counts are printed below observed counts  
Chi-Square contributions are printed below expected counts

	Pos	NotPos	Total
1	16	9	25
	17.35	7.65	
	0.105	0.237	
2	18	6	24
	16.65	7.35	
	0.109	0.247	
Total	34	15	49

Chi-Sq = 0.698, DF = 1, P-Value = 0.404

