L2. Numerical Measures of Data (Cont'd) - Relative Standing

Example 1: Jelly Beans

Jelly Belly is basically the mad scientist of the candy world, whipping up over 100 flavors—some delightful, some downright questionable. The all-time MVPs? Very Cherry, Buttered Popcorn, and Juicy Pear. But not every bean gets to live the sweet life forever. Retired flavors, affectionately (or tragically) known as "has-beans," include Buttered Toast, Draft Beer, Moldy Cheese, and Toothpaste.

At a contest, contestant were asked to guess, how many jelly beans were in a kilogram. Their answers are shown below:

852	861	862	862	863	865	868	871	878	880
882	882	888	891	891	894	894	895	898	899
899	903	907	909	919	926	926	934	937	948

a. Calculate the median.

b. Calculate the 25^{th} percentile.

c. Calculate the 75^{th} percentile.

d. Calculate the 55^{th} percentile.

e. Calculate the 30^{th} percentile.

Solution

a. Median:

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)50}{100} = 15.5 \quad \Rightarrow \quad \begin{aligned} P_{50} &= 891 + 0.5(894 - 891) \\ &= 892.5 \end{aligned}$$

b. 25^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)25}{100} = 7.75 \quad \Rightarrow \quad \begin{aligned} Q_1 &= 868 + 0.25(871 - 868) \\ &= 870.25 \end{aligned}$$

c. 75^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)75}{100} = 23.25 \quad \Rightarrow \quad \begin{aligned} Q_3 &= 907 + 0.75(909 - 907) \\ &= 907.50 \end{aligned}$$

d. 55^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)55}{100} = 17.05 \quad \Rightarrow \quad \begin{aligned} P_{55} &= 894 + 0.05(895 - 894) \\ &= 894.05 \end{aligned}$$

e. 30^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)30}{100} = 9.3 \quad \Rightarrow \quad \begin{aligned} P_{30} &= 878 + 0.3(880 - 878) \\ &= 878.6 \end{aligned}$$

Example 2: Grading

Academic marking was invented in Cambridge in 1792 - until then no one had thought to assign a numerical value to a piece of work. The grades (out of 100) for the first Class Exercise in this course are shown below:

21	22	22	25	25	25	25	26	32	32
34	37	41	42	42	46	49	59	63	67
71	72	73	73	74	77	78	82	82	83
83	89	92	92	94	95	95	97	98	100

a. Calculate the median and interpret its meaning in the context of the problem.

b. Calculate the interquartile range (IQR) and interpret it in the context of the problem.

c. Calculate the 99^{th} percentile.

d. Calculate the 80^{th} percentile.

e. Calculate the 90^{th} percentile.

f. Calculate the 20^{th} percentile.

Solution

a. Median:

$$L = \frac{(N+1)P_i}{100} = \frac{(40+1)50}{100} = 20.50 \quad \Rightarrow \quad P_{50} = 67 + 0.5(71-67)) = 69$$

Interpretation: Half of the class scored 69% or lower on the first assignment, while the other half scored 69% or higher.

b. Interquartile range:

$$L = \frac{(N+1)P_i}{100} = \frac{(40+1)25}{100} = 10.25 \quad \Rightarrow \quad Q_1 = 32 + 0.25(34-32) = 32.5$$

$$L = \frac{(N+1)P_i}{100} = \frac{(40+1)75}{100} = 30.75 \quad \Rightarrow \quad Q_3 = 83$$

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 83 - 32.5 \\ &= 50.5 \end{aligned}$$

Interpretation: The middle 50% of grades are spread over a range of 50.5 percentage points.

To assess if the IQR is large or small, we compute the *Coefficient of Quartile Variation*

$$CQV = \frac{IQR}{Q_1 + Q_3} = \frac{50.5}{32.5 + 83} = 0.4372$$

The CQV is moderately high. This suggests a high variability in student performance: some students are doing very well, while others are struggling.

Remark: $CQV > 0.5 \quad \Rightarrow \quad$ high variability

$CQV < 0.25 \quad \Rightarrow \quad$ low variability

Example 3: The Horlicks Mountains

Horlicks, the malted milk drink best known for knocking people out before bedtime, once had a much grander legacy. In 1930, explorer Richard Byrd, powered by Horlicks (and probably a lot of frostbite), decided to name an entire Antarctic mountain range "Horlick Mountains" in honor of the company that kept his expedition both hydrated and funded during his trek to the South Pole.

Several people were asked how many cups of Horlicks they drank in the past week to help them get to sleep. There responses are shown below:

Cups of Horlicks	Number of People
2	12
4	17
6	15
8	21
10	19
12	16
14	20

a. Calculate the median and interpret its meaning in the context of the problem.

b. Calculate the first quartile.

c. Calculate the third quartile.

d. Calculate the interquartile range and interpret it in the context of the problem.

e. Calculate the 98^{th} percentile.

f. Calculate the 60^{th} percentile.

g. Calculate the 5^{th} percentile.

Solution

Here is the table with added columns to facilitate our calculations.

Cups of Horlicks	Number of People	$LTCF$
2	12	12
4	17	29
6	15	44
8	21	65
10	19	84
12	16	100
14	20	120

a. Median

$$L = \frac{(N+1)P_i}{100} = \frac{(120+1)50}{100} = 60.50 \quad \Rightarrow \quad P_{50} = 8 \text{ cups}$$

Interpretation: 50% of the respondents consumed fewer than 8 cups of Horlicks last week, while the remaining 50% of the respondents drank more than 8 cups of Horlicks last week.

b. First Quartile

$$L = \frac{(N+1)P_i}{100} = \frac{(120+1)25}{100} = 30.25 \quad \Rightarrow \quad Q_1 = 6 \text{ cups}$$

c. Third Quartile

$$L = \frac{(N+1)P_i}{100} = \frac{(120+1)75}{100} = 90.75 \quad \Rightarrow \quad Q_3 = 12 \text{ cups}$$

d. Interquartile range

$$IQR = Q_3 - Q_1 = 12 - 6 = 6 \text{ cups}$$

Interpretation: The middle 50% of the respondents drank 6 cups of Horlicks last week. To assess the magnitude of the IQR, we compute the CVQ

$$CQV = \frac{IQR}{Q_1 + Q_3} = \frac{6}{12+6} = 0.4444$$

A value of 0.444 indicates that the middle 50% of the respondents had a reasonably wide range of Horlicks consumption within the last week.

e. 98^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(120+1)98}{100} = 118.58 \quad \Rightarrow \quad P_{98} = 14 \text{ cups}$$

f. 60^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(120+1)60}{100} = 72.6 \quad \Rightarrow \quad P_{60} = 10 \text{ cups}$$

g. 5^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(120+1)5}{100} = 6.05 \quad \Rightarrow \quad P_5 = 2 \text{ cups}$$

Example 4: Tortoises

In 2011, a celebrity pair of Galápagos tortoises broke up in spectacular fashion. Though the couple had happily cohabitated for about 90 years, they suddenly could no longer stand each other. They now live on opposite sides of a glass wall and Bibi hisses whenever she spots her ex.

Galápagos tortoises can weigh between 135 kg to 400 kg depending on their age, gender, and whether they are living in their natural habitat or in captivity. The table below shows the weights of several Galápagos tortoises:

Weight (kg)	Number of Tortoises
135 – 164	19
165 – 194	16
195 – 224	36
225 – 254	44
255 – 284	22
285 – 314	21
315 – 344	11
345 – 374	20
375 – 404	22

a. Calculate the 90^{th} percentile and interpret its meaning in the context of the problem.

b. Calculate the median and interpret its meaning in the context of the problem.

c. Calculate the first quartile.

d. Calculate the third quartile.

e. Calculate the 40^{th} percentile.

f. Calculate the 15^{th} percentile.

Solution

Here is the table augmented with additional columns.

Weight (kg)	Number of Tortoises	m_i	$LTCF$
135 – 164	19	149.5	19
165 – 194	16	179.5	35
195 – 224	36	209.5	71
225 – 254	44	239.5	113
255 – 284	22	269.5	137
285 – 314	21	299.5	158
315 – 344	11	329.5	169
345 – 374	20	359.5	189
375 – 404	22	389.5	211

a. 90^{th} percentile

$$L = \frac{(N+1)P_i}{100} = \frac{(211+1)90}{100} = 190.8 \quad \Rightarrow \quad P_{90} = 389.5\,kg$$

Interpretation: 90% of Galápagos tortoises weigh less than 389.5 kg, while 10% of Galápagos tortoises weigh more than 389.5 kg.

b. Median

$$L = \frac{(N+1)P_i}{100} = \frac{(211+1)50}{100} = 106 \quad \Rightarrow \quad P_{50} = 239.5\,kg$$

Interpretation: 50% of Galápagos tortoises weigh less than 239.5 kg, while the remaining 50% of Galápagos tortoises weigh more than 239.5 kg.

c. First quartile.

$$L = \frac{(N+1)P_i}{100} = \frac{(211+1)25}{100} = 53 \quad \Rightarrow \quad Q_1 = 209.5\,kg$$

d. Third quartile.

$$L = \frac{(N+1)P_i}{100} = \frac{(211+1)75}{100} = 159 \quad \Rightarrow \quad Q_3 = 329.5\,kg$$

e. 40^{th} percentile.

$$L = \frac{(N+1)P_i}{100} = \frac{(211+1)40}{100} = 848 \quad \Rightarrow \quad P_{40} = 239.5\,kg$$

f. 15^{th} percentile.

$$L = \frac{(N+1)P_i}{100} = \frac{(211+1)15}{100} = 31.8 \quad \Rightarrow \quad P_{15} = 149.5\,kg$$

Example 5: Men in Grey Suits

'Men in grey suits' is Australian slang for sharks. Surfers use it to alert people to their presence without causing too much anxiety. The lifespans of several male and female great white sharks are shown below:

Male Great White Sharks:

24	30	31	31	31	32	33	33	33	33
34	34	34	36	37	38	38	39	39	40
40	41	41	42	43	43	43	44	45	47

Female Great White Sharks:

30	31	31	32	33	35	35	36	36	36
36	38	39	39	39	39	39	41	41	41
42	43	43	43	43	44	45	52	53	55

a. Compute and interpret the IQR for the male great white shark data.

b. Identify any outlier in the male great white shark data if they exist.

c. Compute and interpret the IQR for the female great white shark data.

d. Identify any outlier in the female great white shark data if they exist.

e. Compare and contrast the two datasets.

Solution

a. IQR for male shark data.

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)25}{100} = 7.75 \quad \Rightarrow \quad Q_1 = 33$$

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)75}{100} = 23.25 \quad \Rightarrow \quad Q_3 = 41 + 0.25(42-41)$$
$$= 41.25$$

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 41.25 - 33 \\ &= 8.25 \end{aligned}$$

Interpretation: An IQR of 8.25 years for the male shark data indicates that the middle 50% of male sharks have ages spanning 8.25 years.

$$CVQ = \frac{IQR}{Q_1 + Q_3} = \frac{8.25}{33 + 41.25} = 0.006$$

A CQV of 0.006 for the male shark data indicates that there is extremely low variability in the middle 50% of ages, meaning most male sharks have nearly identical ages with minimal dispersion.

b. Outlier detection

$$\begin{aligned} Q_1 - 1.5(IQR) &= 33 - 1.5(8.25) \\ &= 20.625\, kg \end{aligned}$$

$$\begin{aligned} Q_3 + 1.5(IQR) &= 41.25 + 1.5(8.25) \\ &= 53.625\, kg \end{aligned}$$

Conclusion: $\because$ there are no values in the male shark dataset that is smaller than 20.625, nor are there any data points greater than 53.625 $\Rightarrow$ there are no outliers in the data set.

c. IQR for female shark data.

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)25}{100} = 7.75 \quad \Rightarrow \quad Q_1 = 35 + 0.75(36 - 35) = 35.75$$

$$L = \frac{(N+1)P_i}{100} = \frac{(30+1)75}{100} = 23.25 \quad \Rightarrow \quad Q_3 = 43$$

$$\begin{aligned} IQR &= Q_3 - Q_1 \\ &= 43 - 35.75 \\ &= 7.75 \end{aligned}$$

Interpretation: An IQR of 7.75 years for the female shark data indicates that the middle 50% of female sharks have ages spanning 7.75 years.

$$CVQ = \frac{IQR}{Q_1 + Q_3} = \frac{7.75}{35.75 + 43} = 0.0984$$

A CQV of 0.0984 for the female shark data indicates that there is low variability in the middle 50% of ages, meaning most female sharks have relatively similar ages with only slight dispersion.

d. Outlier detection

$$\begin{aligned} Q_1 - 1.5(IQR) &= 35.75 - 1.5(7.75) \\ &= 24.125\, kg \end{aligned}$$

$$\begin{aligned} Q_3 + 1.5(IQR) &= 43 + 1.5(7.75) \\ &= 54.625\, kg \end{aligned}$$

Conclusion: There is one data value which is greater than $54.625\,kg$. Thus, 55 is an outlier.

e. Comparison of datasets.

The median lifespan of male sharks (37.5 years) is slightly lower than that of female sharks (39 years). This trend is also reflected in the range of each dataset, with males spanning 23 years ($47 - 24$) and females spanning 25 years ($55 - 30$). The CQV further highlights this difference, showing very little variability in the middle 50% of male shark ages (0.0006) compared to the low variability in female shark ages (0.0984). Since the male data has no outliers, the mean and standard deviation serve as reliable measures of central tendency and spread. However, for the female data, the presence of an outlier makes the median and interquartile range more appropriate, as the mean and standard deviation could be unduly influenced by the extreme value.