Section 3.3: Measures of Position

1 z-scores

We've used the empirical rule to examine distributions of data in the normal curve and compare some values. To formalize the process we introduce the z-score function. We use a z-score to measure the distance a value is from the mean relative to standard deviation.

$$z = \frac{x - \mu}{\sigma} \text{ for populations}$$

$$z = \frac{x - \bar{x}}{s} \text{ for samples}$$

Think of standard deviation as a type of ruler that has no units of measurement.

Problem 1 *Consider Math 1107/01, with a test 1 average of 10 and standard deviation of 2 and Math 1107/02 with a test 1 average of 150 and standard deviation of 15. Which score is better, a 14 from Math 1107/01 or a 160 from Math 1107/02?*

Problem 2 *Par on the Jurassic Mini Golf Course (http://myrtlebeachfamilygolf.com/jurassic-golf/) in Myrtle Beach, SC is 44 strokes with a standard deviation of 8. Par on the Captain Kidd's Challenge (http://www.piratesislandgolf.com/) in Hilton Head, SC is 56 strokes with a standard deviation of 12.*

1. Which score is better, a 42 at Jurassic Golf or a 60 at Captain Kidd's Challenge?

2. Which score is better, a 50 at Jurassic Golf or a 60 at Captain Kidd's Challenge?

Problem 3 *Use z-scores to determine which score is better, Evan who scored a 12 in Math 1107/01 or Francine who scored a 160 in Math 1107/02?*

Problem 4 *Determine which score is better, a 14.5 in Math 1107/01 or a 133 in Math 1107/02? Do we really need to use a z-score for this problem?*

Remark 5 *The sign on a z-score is important! A negative z-score tells us that the data value is below the mean, while a positive z-score tells us that the data value is above the mean.*

Example 6 *What is the original test score for a z-score of* -1.5 *in Math 1107/01? We solve* $-1.5 = \frac{x-10}{2}$ *for* x *and find the test score* $x = 7$

Problem 7 *What is the original test score for a z-score of* -2 *in Math 1107/02?*

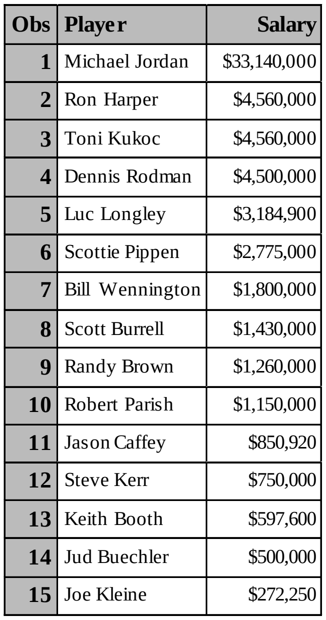
2 Percentiles

As noted earlier, for every median, 50% of the data falls below the median and 50% falls above the median. Let's extend this notion for values other than 50%.

Definition 8 *For a given set of data, the* ***pth percentile*** *is a number x such that p% of the data falls below x. Consequently, (100-p)% falls above x.*

Another name for the median is P_{50}, the 50th percentile. Two other very common percentile scores with special names are $Q_1 = P_{25}$, the lower quartile and $Q_3 = P_{75}$, the upper quartile. Sometimes the median is referred to as Q_2.

Example 9 *Which Chicago Bulls salary would you prefer to be paid* P_{10} *or* P_{80}*? Explain. Since* $P_{10} = \$539,040$ *while* $P_{80} = \$4,512,000$*, I personally would prefer* P_{80} *as a salary. Don't confuse the top 10% of the data with* P_{10}*! Percentile scores are always based on the percentage of data that falls below!*



Obs	Player	Salary
1	Michael Jordan	$33,140,000
2	Ron Harper	$4,560,000
3	Toni Kukoc	$4,560,000
4	Dennis Rodman	$4,500,000
5	Luc Longley	$3,184,900
6	Scottie Pippen	$2,775,000
7	Bill Wennington	$1,800,000
8	Scott Burrell	$1,430,000
9	Randy Brown	$1,260,000
10	Robert Parish	$1,150,000
11	Jason Caffey	$850,920
12	Steve Kerr	$750,000
13	Keith Booth	$597,600
14	Jud Buechler	$500,000
15	Joe Kleine	$272,250

Chicago Bulls Salaries 1997-1998 Season

Let's examine $P_{10} = \$539,040$. Since there are 15 data points in the set of player salaries, technically P_{10} should be greater than exactly 1.5 of the data points. Obviously that is impossible since you cannot have half a piece of data. This problem is due to the fact that our collection of data is quite small. Small data sets cause a myriad of difficulties as we shall see throughout the course. Watch out! Different technologies may compute quartiles and percentiles in slightly different ways. For large populations with a normal distribution, it is quite easy to compute percentile scores.

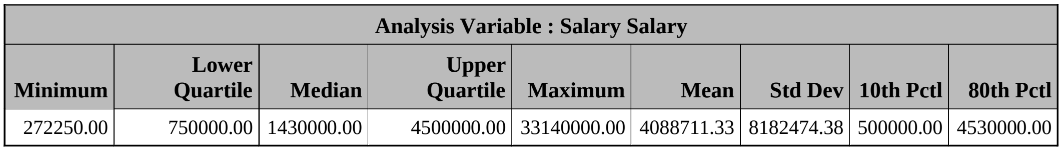
3 The Five Number Summary

Definition 10 *For any set of data the* ***five number summary*** *is, in order, the five summary statistics:*

$$minimum, Q_1, median, Q_3, maximum.$$

The five number summary for the player salaries for the 1997-1998 Chicago Bulls is

	Excel	TI-83/84
minimum	\$272,250	\$272,250
Q_1	\$800,460	\$750,000
median	\$1,430,000	\$1,430,000
Q_3	\$3,842,450	\$4,500,000
maximum	\$33,140,000	\$33,140,000



Analysis Variable : Salary Salary								
Minimum	Lower Quartile	Median	Upper Quartile	Maximum	Mean	Std Dev	10th Pctl	80th Pctl
272250.00	750000.00	1430000.00	4500000.00	33140000.00	4088711.33	8182474.38	500000.00	4530000.00

SAS output

4 IQR

The range is very susceptible to unusually large or small values in a date set. A single extreme value skews the range of a set of data. The interquartile range (IQR) is much more resistant to skew. The IQR measures the range of the central 50% of the data.

Definition 11 *For a given set of data, the **IQR** is the (positive or occasionally 0) difference between Q_3 and Q_1 in a quantitative data set.*

Example 12 *For the player salaries of the 1997-1998 Chicago Bulls the range is* $33140000 - 272250 = 32\,867\,750$ *and the* $IQR = 3842450 - 800460 = 3041\,990$.

5 Outliers and IQR

We consider values that are unusually large or small compared to the rest of a data set to be **outliers**. We formally define unusually large or small via the IQR.

$$\text{lower outlier boundary } = Q_1 - 1.5 * IQR$$

$$\text{upper outlier boundary } = Q_3 + 1.5 * IQR$$

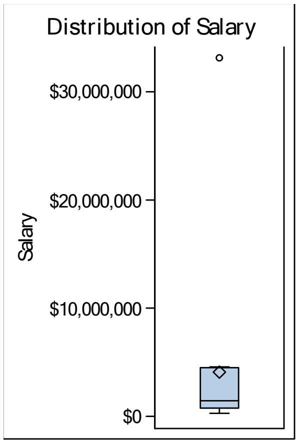
Values that fall outside these boundaries are outliers.

Problem 13 *Determine the salary outliers for the '97-'98 Bulls roster.*

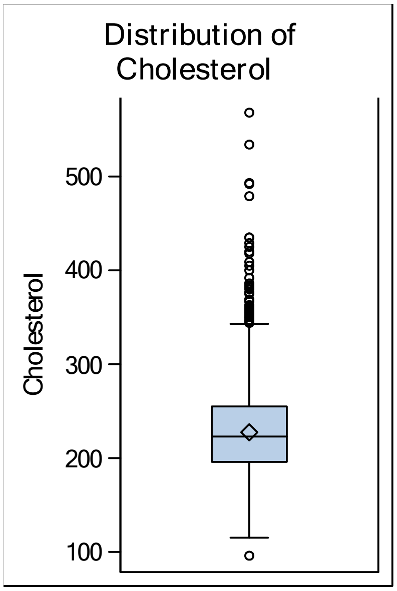
We've already determined that $Q_1 = \$800,460$, $Q_3 = \$3,842,450$ and $IQR = 3842450 - 800460 = 3041\,990$. So, the lower outlier boundary is $800460 - 1.5 * 3041\,990 = -3762500$. Clearly, nothing falls below this lower outlier boundary. The upper outlier boundary is $3842450 + 1.5 * 3041\,990 = 8405400$. Only Jordan's salary is larger than $8,405,400. The only outlier is this data set is the salary for Michael Jordan.

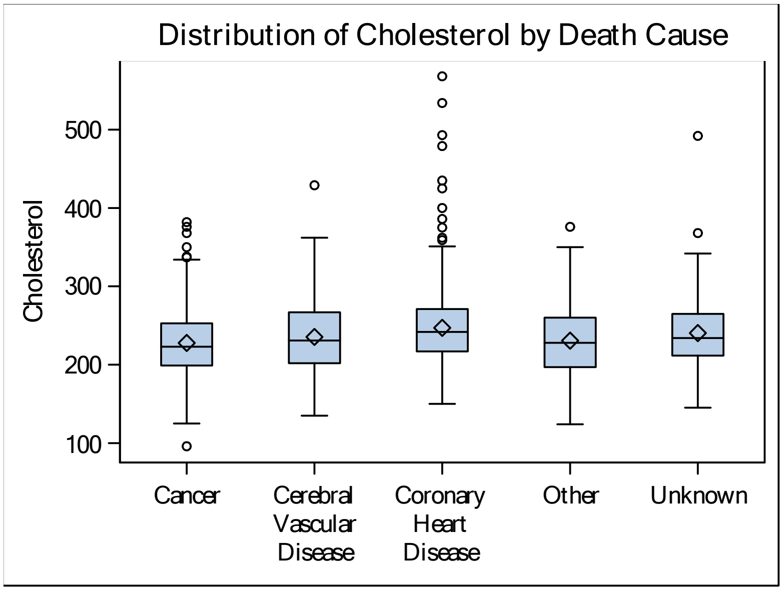
6 Box and Whisker Plots

A box and whisker plot is a graphical representation of Q_1, Q_3, the mean, the median and outlier boundaries. The bottom and top of the box represent Q_1 and Q_3. The line inside the box is the median value. The mark inside the box represents the mean. The whiskers represent the extreme values that are not outliers. Points outside these whiskers are outliers.



The box and whisker plot for the '97-'98 Chicago Bull's salaries





7 Exercises

1. Navidi/Monk Section 3.3: 5-8, 13, 14, 17, 18, 21-24, 26a-e, 27a-e, 38a-c