Section 3.2: Measures of Variability

The mean and median are good statistics to employ when describing the center of a collection of data. However, there is more to a collection of data than just the center! Recall our example of average test scores in two different classes.

Example 1 *The average score on Test 1 in MATH 8000 at the University of Nowhere is 75. Of the 100 students in the class, half scored a 50 and the other half scored 100.*

Example 2 *The average score on Test 1 in MATH 8000 at the University of Nowhere is 75. Each of the 100 students in the class scored a 75.*

Problem 3 *What is the median score in each class?*

Knowing the mean and median is a good start to understanding data. But it is not enough. We must also understand how data varies. The same unit of measurement may not always have the same value or meaning.

Example 4 *Consider two teams of five cross country runners. Both teams average a 9 minute mile. Who wins a mile race? Consider the mile times (in minutes) for each team member given in the following table.*

Team 1	*Alice*	*Bob*	*Chris*	*David*	*Emily*
mile time	*9*	*9*	*9*	*9*	*9*
Team 2	*Frank*	*Greg*	*Hannah*	*Ian*	*Jenny*
mile time	*11*	*8*	*11*	*8*	*7*

While both teams average a 9 minute mile, Jenny wins the race for her team. Knowing the center of a collection of data is important but there is also the need to understand how the data varies.

1 Range

The simplest measure of the **spread** (**dispersion** or **variability**) of data is the range.

Definition 5 *For a given set of data, the **range** is the (positive or occasionally 0) difference between the largest and smallest values in a quantitative data set.*

Example 6 *The range of mile times for team 1 is* $9 - 9 = 0$. *The range of mile times for team 2 is* $11 - 7 = 4$.

Problem 7 *What is the range of salaries for the Chicago Bull's '97-'98 roster?*

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

Problem 8 *Is "range" a resistant function?*

2 Variance and Standard Deviation

The most important and commonly used measure of spread is the standard deviation.

Definition 9 *Standard deviation measures the spread of the data from the mean. This can be seen at the heart of the formula for standard deviation. We denote a sample standard deviation by s and a population standard deviation by σ. There is a subtle difference in the two formulae.*

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}}$$

Definition 10 *Variance for samples and population is s^2 and σ^2.*

Remark 11 *Note that as n grows larger and larger the difference between s and σ grows smaller and smaller. This is not strictly a limit problem as out sets are finite but the obvious similarity is present.*

Know how to compute variance and standard deviation on the TI 83/84. For the player salaries on the 1997-1998 Chicago Bulls the sample standard deviation is \$8,182,474.38 and the population standard deviation is.\$7,905,021.27.

Example 12 *What is the sample variance for Bull's salaries? Variance is always standard deviation squared. So, sample variance for Bull's salaries is* $s^2 = 8182474.38^2 = 6.695\,3 \times 10^{13}$.

2.1 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 13 *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 14 *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

3 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 15 *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 16 *For the player salaries of the 1997-1998 Chicago Bulls the range is* $33140000 - 272250 = 32\,867\,750$ *and the* $IQR = 3842450 - 800460 = 3041\,990$.

Problem 17 *Find* Q_1, Q_3 *and* IQR *for the Chicago Bull's salaries.*

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 Using R to compute measures of spread

First item of note is that R has a single command for variance and standard deviation. These commands compute sample variance and sample standard deviation.

```
> list = c(2, 4, 4, 4, 5, 5, 7, 9, 10, 57)
> var(list)
[1] 270.6778
> sd(list)
[1] 16.45229
> range(list)
[1]  2 57
> fivenum(list)
[1]  2  4  5  9 57
> IQR(list)
[1] 4.5
> fivenum(bulls_salaries$Salary)
[1]   272000   800460  1430000  3842450 33140000
> quantile(list,probs = seq(0, 1, 1/4))
  0%  25%  50%  75% 100%
 2.0  4.0  5.0  8.5 57.0
> quantile(bulls_salaries$Salary,probs = seq(0, 1, 1/4))
      0%      25%      50%      75%     100%
  272000   800460  1430000  3842450 33140000
> quantile(bulls_salaries$Salary,probs = seq(0, 1, 1/5))
      0%      20%      40%      60%      80%     100%
  272000   719520  1216000  2190000  4512000 33140000
```

5 Exercises

1. Kokoska 3rd edition Section 3.2: 3.41 a, c, 3.44 a, c, 3.45, 3.47

2. We have computed the mean, median and standard deviation for the 1997-1998 Chicago Bulls salaries. Suppose that every player receives a $1,000,000 raise? Find the values for min, max, mean, median, standard deviation, Q_1, Q_3 and IQR for the post raise salaries. How have these values changed?

3. We have computed the mean, median and standard deviation for the 1997-1998 Chicago Bulls salaries. Suppose that every player receives a 10% raise? Find the values for min, max, mean, median, standard deviation, Q_1, Q_3 and IQR for the post raise salaries. How have these values changed?

4. Let S_1 be a data set whose mean value is m. Let S_2 be the resulting set when the constant k is added to every value in S_1. Prove that the mean of S_2 is $m + k$.

5. Let S_1 be a data set whose mean value is m. Let S_2 be the resulting set when every value in S_1 is multiplied by the constant k. Prove that the mean of S_2 is km.

6. Let's play a game! Every student gets to play this game once. I have two boxes up front on my gaming table. A single play of this game consists of a student selecting a box and then randomly selecting a ticket from the box. The student then receives the value of that ticket.

Box A

\$0	\$4	\$500	\$997
\$1	\$5	\$994	\$998
\$2	\$6	\$995	\$999
\$3	\$500	\$996	\$1000

Box B

\$0	\$500	\$500	\$500
\$500	\$500	\$500	\$500
\$500	\$500	\$500	\$500
\$500	\$500	\$500	\$1000

Which box will you pick? There is no single correct answer but you should be prepared to defend your selection. Compute mean, median and standard deviation and use some or all of these values to back up your answer.