Section 1.2: Summary Statistics

In statistics we frequently need to perform operations on entire sets of data. One such operation is to add all values in a particular set together. We use a capital Greek sigma, Σ, to tell us to add everything together. So $\sum$Chris' test scores is $79 + 88 + 92 = 259$.

MATH 1107 Chris	Tests	Quizzes
	79	95
	88	86
	92	75
		89
		81

If the data set in question is obvious we may just write $\sum x$ where x represents every value is the data set. The sum of Chris' quiz scores is $\sum x = 95 + 86 + 75 + 89 + 81 = 426$.

Example 1 *For Chris' test scores,* $\sum x^2 = 79^2 + 88^2 + 92^2 = 22\,449$. *In contrast,* $\left(\sum x\right)^2 = (79 + 88 + 92)^2 = 259^2 = 67\,081$.

1 The Mean

Definition 1 *The* ***mean*** *(also known as the* ***average****) in a set of n observations, $x_1, x_2, ..., x_n$, is*

$$\frac{\sum x_i}{n}. \tag{1}$$

Notation 1 *The mean is always calculated as the sum of all observations divided by the number of observations. However, when we practice inferential statistics, it is vital to indicate if the average comes from a sample or a population. We denote a sample average by $\overline{x}$ and a population average by μ.*

Example 2 *The mean test score for Chris is* $\mu = \dfrac{\sum x_i}{n} = \dfrac{79 + 88 + 92}{3} = \dfrac{259}{3} = 86.333$

Problem 2 *Find the mean quiz score for Chris.*

2 The Median

Definition 2 *The* ***median*** *in a set of n observations that are ordered from smallest to largest is the middle observation (if n is odd) or the mean of the two middle observations (if n is even).*

Example 3 *To determine the median quiz score by Chris we first order the data from smallest to largest:*

$$75, 81, 86, 89, 95$$

Now it is clear that the middle observation, and hence the median, is 86.

In the case of Chris' quiz scores the median and the mean are pretty close to each other. That won't always be the case!

Example 4 *In CHEM 1101, Chris took six quizzes. His scores are: 81, 76, 82, 78, 12, 75. Again, we order the scores from smallest to largest.*

$$12, 75, 76, 78, 81, 82$$

Here we average the two middle most observations and the median is $\frac{76+78}{2} = 77$*. Just for fun, Chris' average quiz score in CHEM 1101 is* $\mu = \frac{\sum x_i}{n} = \frac{12+75+76+78+81+82}{6} = \frac{202}{3} = 67.333$*.*

Problem 3 *Here the median and mean are not very close to each other. Why?*

Definition 3 *A statistic is* ***resistant*** *if its value is not susceptible to extreme data points.*

Take note that the median cuts the data set in half. That is, 50% of the data in the set falls below the median and 50% of the date falls above the median. This is not necessarily true about the average.

Example 5 *Consider the player salaries for the 1997-1998 Chicago Bulls roster. Here the average salary is almost three times the median salary. This is due to Michael Jordan's salary which is significantly larger (though he was still underpaid!) than all other salaries.*

	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
	Average	$4,088,711
	Median	$1,430,000

Chicago Bulls Salaries
1997-1998 Season

Remark 4 *The median is resistant. The mean is not resistant.*

3 The Mode

Definition 4 *The **mode** in a set of n observations is the value that occurs most frequently. A data set may have more than one mode. A data set with two modes is called **bimodal**. A data set with more than two modes is called **multi-modal**.*

Example 6 *There is no mode for Chris for either test scores or quiz scores. There is no value that appears more frequently than any other.*

Example 7 *The mode of number of games played for the 2012 Atlanta Hawks is 77. The mode for games started is 0.*

Totals Glossary · SHARE · Embed · CSV · PRE · LINK · ?

Rk	Player	Age	G	GS	MP	FG	FGA	FG%	3P	3PA	3P%	FT	FTA	FT%	ORB	DRB	TRB	AST	STL	BLK	TOV	PF	PTS
1	Joe Johnson	29	72	72	2554	514	1161	.443	89	300	.297	195	243	.802	59	232	291	338	47	7	146	131	1312
2	Josh Smith	25	77	77	2645	497	1041	.477	51	154	.331	229	316	.725	134	523	657	255	99	120	197	217	1274
3	Al Horford	24	77	77	2704	513	921	.557	2	4	.500	150	188	.798	182	536	718	266	59	80	119	193	1178
4	Jamal Crawford	30	76	0	2297	368	874	.421	119	349	.341	222	260	.854	22	108	130	241	57	14	145	97	1077
5	Marvin Williams	24	65	52	1865	246	537	.458	37	110	.336	147	174	.845	68	245	313	88	34	23	62	104	676
6	Mike Bibby	32	56	56	1674	192	441	.435	113	256	.441	29	46	.630	15	128	143	202	38	6	68	125	526
7	Jeff Teague	22	70	7	963	133	304	.438	18	48	.375	77	97	.794	11	91	102	138	45	25	64	82	361
8	Zaza Pachulia	26	79	7	1244	107	232	.461	0	0		135	179	.754	119	214	333	58	34	22	69	184	349
9	Josh Powell	28	54	0	653	94	208	.452	0	1	.000	36	45	.800	49	86	135	22	5	5	53	78	224
10	Maurice Evans	32	47	12	837	79	201	.393	28	89	.315	24	28	.857	23	61	84	30	16	5	15	74	210
11	Kirk Hinrich	30	24	22	686	80	185	.432	32	76	.421	14	21	.667	7	46	53	78	19	7	37	66	206
12	Damien Wilkins	31	52	0	676	69	137	.504	2	10	.200	40	56	.714	23	67	90	41	27	9	21	72	180
13	Jason Collins	32	49	28	593	34	71	.479	1	1	1.000	27	41	.659	30	72	102	22	9	9	26	97	96
14	Jordan Crawford	22	16	0	160	27	77	.351	9	27	.333	4	6	.667	9	19	28	15	3	0	15	13	67
15	Etan Thomas	32	13	0	82	10	21	.476	0	0		12	15	.800	6	17	23	2	1	4	5	11	32
16	Hilton Armstrong	26	12	0	76	6	12	.500	1	1	1.000	2	10	.200	3	14	17	4	3	5	3	9	15
17	Pape Sy	22	3	0	21	2	6	.333	0	1	.000	3	3	1.000	2	1	3	2	1	0	3	1	7

2011 Atlanta Hawks Team Statistics

One cannot compute a mean or median for qualitative data. It is possible to compute the mode for qualitative data.

Problem 5 *A small bag of M&M's contained the following 20 candies. Find the mode for color of M&M.*

red	green	blue	red	brown
red	brown	brown	orange	green
orange	orange	green	green	blue
green	green	brown	red	red

4 Trimmed Means

Definition 5 *The p% trimmed mean is found by first trimming (removing) the top p% and bottom p% of the data set. Second, compute the mean of the remaining values.*

Example 8 *Let* $S = \{3, 4, 4, 5, 6, 6, 7, 8, 15, 30\}$ *The mean of* S *is* $\frac{3+4+4+5+6+6+7+8+15+30}{10} = 8.8$*. The 20% trimmed mean of* S *is* $\frac{4+5+6+6+7+8}{6} = 6$*. Note that the trimmed mean removes outliers.*

5 The Proportion of Success

Definition 6 *For a data set where each outcome can be classified as a success or failure, the* ***proportion of success*** *is a measure of center. The proportion of success is computed by the number of successes in the data set divided by the number of observations.*

Example 9 *A single die is rolled 10 times. The sequence of rolls is* $(1,4,5,6,3,4,5,2,3,2)$. *If a success is rolling an odd prime number then the proportion of success is* $p = \frac{4}{10} = 0.4$.

Problem 6 *If a success is rolling a perfect square, what is the proportion of success for the above sequence?*

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 10 *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 11 *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 7 *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 12 *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.

6 Range

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

Definition 7 *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 13 *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 8 *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 9 *Is "range" a resistant function?*

7 Variance and Standard Deviation

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

Definition 8 *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 9 *Variance for samples and population is s^2 and σ^2.*

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

7.1 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 10 *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 15 *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

8 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 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 10 *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

9 Exercises

1. Navidi Section 1.2: 1-9, 10a, c, d, e, 16
2. Compute the 20% trimmed mean for the '97-'98 Chicago Bulls salaries.
3. What difficulty do you encounter if you wish to compute a 10% trimmed mean for the '97-'98 Chicago Bulls salaries.
4. If possible, create a set of 10 integer data points between 0 and 10 such that 90% of the data is less than the mean. Your data points need not be unique. If not posssible, explain why.
5. If possible, create a set of 10 integer data points between 0 and 10 such that 90% of the data is less than the median. Your data points need not be unique. If not posssible, explain why.
6. Is it possible for John to lead the NFL in total rushing yards but for Nick to lead in average rushing yards per game played in the same season? If yes, construct a set of data that demonstrates it is possible. If no, explain why not. **HINT!** It is possible.