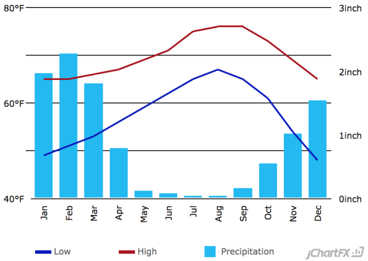
Where would you rather live? (And why?)

CityA

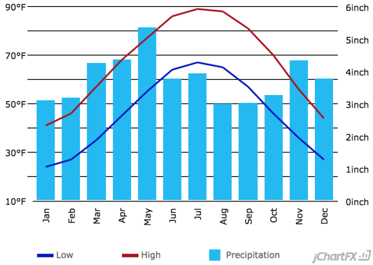


	Jan	Feb	Mar	Apr	May	Jun
Average high in °F:	65	65	66	67	69	71

	Jul	Aug	Sep	Oct	Nov	Dec
Average high in °F:	75	76	76	73	69	65

Annual high temperature: 69.8°F

CityB



	Jan	Feb	Mar	Apr	May	Jun
Average high in °F:	41	46	57	68	77	86

	Jul	Aug	Sep	Oct	Nov	Dec
Average high in °F:	89	88	81	70	56	44

Annual high temperature: 66.9°F

City A is San Diego, CA, and City B is Evansville, IN

Measures of Dispersion

Suppose you need a new quarterback for your football team and you are trying to decide between two quarterbacks who have played in the same league last season with roughly the same strength of schedule. The number of completions in each game for the 16 games of the previous season are shown for each quarterback:

Game	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Quarterback A:	27	30	32	28	33	28	29	30	25	24	19	22	37	27	32	25
Quarterback B:	25	33	40	19	39	35	17	12	32	33	12	39	30	17	35	30

Both Quarterbacks have the same average number of completions over the last season, $\mu = 28$. However we see that Quarterback B has a a more varied performance record than Quarterback A. Obviously one needs to take this variability in the data into account when comparing the quarterbacks.

Measures of Dispersion

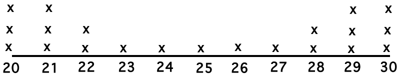
The Range of a set of data is the largest measurement minus the smallest measurement.

Example Calculate the range for the data for Quarterback A and Quarterback B in the example above.

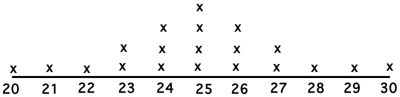
The minimum number of completions for Quarterback A is 19, the maximum is 37. The minimum number of completions for Quarterback B is 12, the maximum is 40. Hence the ranges are 18 for Quarterback A and 28 for Quarterback B.

Measures of Dispersion

Although the range is easy to compute it is a crude measure of variability. Consider the following two sets of data which have the same mean, 25, and the same range, 10, but obvious differences in the pattern of variability:

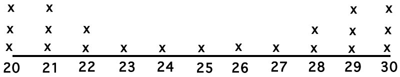


Data Set 1

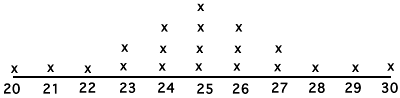


Data Set 2

Measures of Dispersion



Data Set 1



Data Set 2

Data set 1 has most of its values far from the mean and is u-shaped, while data set 2 has most of its values closer to the mean and is mound shaped or bell shaped. In order to catch the different patterns in variability above, we need a more subtle measure than the range. Two widely used measure of consistency (or lack of it) in the data are given by the **variance** and the **standard deviation**. The formula for each depends on whether one is dealing with data from a population or a sample.

Population Variance/standard deviation

For a set of data $\{x_1, x_2, \ldots x_n\}$ for a population of size n, we define the **population variance**, denoted by σ^2, to be the average squared distance from the mean, μ:

$$\boxed{\sigma^2 = \frac{(x_1-\mu)^2 + (x_2-\mu)^2 + \cdots + (x_n-\mu)^2}{n}}$$

As with the calculation of the mean, we can shorten calculations if we have a frequency distribution at our disposal. For **a set of data** for a population of size n, with observed values $\{O_1, O_2, \ldots, O_m\}$ and frequencies $\{f_1, f_2, \ldots, f_m\}$ respectively, the **population variance** is given by:

Population Variance/standard deviation

$$\boxed{\sigma^2 = (O_1 - \mu)^2 \frac{f_1}{n} + (O_2 - \mu)^2 \frac{f_2}{n} + \cdots + (O_m - \mu)^2 \frac{f_m}{n}.}$$

Observations	Deviation	Squared Deviation	Sq. Dev. × Rel. Freq.
O_i	$O_i - \mu$	$(O_i - \mu)^2$	$(O_i - \mu)^2 \frac{f_i}{n}$
O_1 O_2 $\vdots$ O_m	$O_1 - \mu$ $O_2 - \mu$ $\vdots$ $O_m - \mu$	$(O_1 - \mu)^2$ $(O_2 - \mu)^2$ $\vdots$ $(O_m - \mu)^2$	$(O_1 - \mu)^2 \frac{f_1}{n}$ $(O_2 - \mu)^2 \frac{f_2}{n}$ $\vdots$ $(O_m - \mu)^2 \frac{f_m}{n}$
			$\sigma^2 =$ Sum

The **population standard deviation** for the data is the square root of the population variance,

$$\boxed{\sigma = \sqrt{\sigma^2}.}$$

Example 1: Find the variance, σ^2 and standard deviation, $\sqrt{\sigma^2}$, for the number of completions for each Quarterback above. The value of μ is 28 for both players.

Quarterback A

O_i # Completions	f_i Frequency
19	1
22	1
24	1
25	2
27	2
28	2
29	1
30	2
32	2
33	1
37	1

Quarterback B

O_i # Completions	f_i Frequency
12	2
17	2
19	1
25	1
30	2
32	1
33	2
35	2
39	2
40	1

Quarterback A:

Observations	Deviation	Squared Deviation	Sq. Dev. × Rel. Freq.
O_i	$O_i - 28$	$(O_i - 28)^2$	$(O_i - 28)^2 \frac{f_i}{16}$
19	-9	81	$\frac{81 \cdot 1}{16}$
22	-6	36	$\frac{36 \cdot 1}{16}$
24	-4	16	$\frac{16 \cdot 1}{16}$
25	-3	9	$\frac{9 \cdot 2}{16}$
27	-1	1	$\frac{1 \cdot 2}{16}$
28	0	0	$\frac{0 \cdot 2}{16}$
29	1	1	$\frac{1 \cdot 1}{16}$
30	2	4	$\frac{4 \cdot 2}{16}$
32	4	16	$\frac{16 \cdot 2}{16}$
33	5	25	$\frac{25 \cdot 1}{16}$
37	9	81	$\frac{81 \cdot 1}{16}$
			$\sigma^2 = \text{Sum}$

Quarterback A: $\sigma^2 = \dfrac{300}{16} = \dfrac{75}{4} = 18.75 \qquad \sigma \approx 4.3301270189$

Quarterback B:

Observations	Deviation	Squared Deviation	Sq. Dev. × Rel. Freq.
O_i	$O_i - 28$	$(O_i - 28)^2$	$(O_i - 28)^2 \frac{f_i}{16}$
12	-16	256	$\frac{256 \cdot 2}{16}$
17	11	121	$\frac{121 \cdot 2}{16}$
19	-9	81	$\frac{81 \cdot 1}{16}$
25	-3	9	$\frac{9 \cdot 1}{16}$
30	2	4	$\frac{4 \cdot 2}{16}$
32	4	16	$\frac{16 \cdot 1}{16}$
33	5	25	$\frac{25 \cdot 2}{16}$
35	7	49	$\frac{49 \cdot 2}{16}$
39	11	121	$\frac{121 \cdot 2}{16}$
40	12	144	$\frac{144 \cdot 1}{16}$
			$\sigma^2 = \text{Sum}$

Quarterback B: $\sigma^2 = \dfrac{1402}{16} = 87.625 \quad \sigma = 9.3608226134$

Sample Variance/Standard Deviation

If we calculate the variance according to the formula given above, for a sample from a particular population, it is not, on average, accurate as an estimate for the population variance — it on average tends to be too small an estimate. So for a sample from a given population, we use the **sample variance** as an unbiased estimator of the population variance.

Given a sample, $\{x_1, \quad x_2, \quad \ldots \quad x_n\}$, of size n from a population, where the sample mean is given by $\bar{x}$, the **sample variance** is given by

$$\boxed{s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}.}$$

The **sample standard deviation** is given by

$$\boxed{s = \sqrt{s^2}.}$$

If the data is given in a frequency distribution, we can shorten the calculations. If the outcomes in the sample are given by $\{O_1, O_2, \ldots, O_m\}$ with respective frequencies given by $\{f_1, f_2, \ldots, f_m\}$, then

$$\boxed{s^2 = \frac{(O_1 - \bar{x})^2 f_1 + (O_2 - \bar{x})^2 f_2 + \cdots + (O_n - \bar{x})^2 f_m}{n-1}.}$$

Example 3 A random sample of size twenty of a golfer's scores for nine-hole rounds of golf over the past year are as follows:

$$39,\ 40,\ 40,\ 41,\ 39,\ 40,\ 44,\ 43,\ 40,\ 41,\ 40,\ 41,\ ,41,$$

$$42,\ 43,\ 40,\ 41,\ 41,\ 41,\ 43.$$

Compute the mean, sample variance and the sample standard deviation for the sample of the golfer's scores. You can view the sample variance as an estimate of the overall variance of the golfer's scores.

The mean is 41.

Observations	Frequency	Deviation	Squared Deviation	Sq. Dev. × Rel. Freq.
O_i	f_i	$O_i - 41$	$(O_i - 41)^2$	$(O_i - 41)^2 \cdot f_i$
39	2	-2	4	8
40	6	-1	1	6
41	7	0	0	0
42	1	1	1	1
43	3	2	4	12
44	1	3	9	9
Sample size =	20		Sum =	36

Hence $s^2 = \dfrac{36}{20-1} = \dfrac{36}{19} \approx 1.8947368421.$
$s \approx 1.3764944032.$

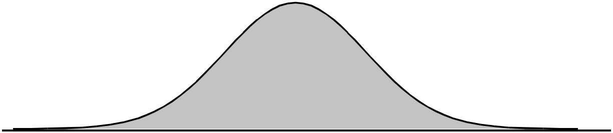
And while we are at it: the median is also 41 and the mode is 41 as well.

Interpreting The Standard Deviation

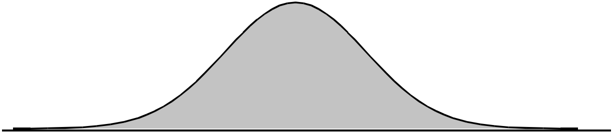
When presented with raw scores for performance, it is difficult to interpret their meaning without some measure of center and variability for the population from which they come. In any set of data, whether it is population data or a sample, observations that are more than 3 standard deviations from the mean are rare and exceptional. One such rule demonstrating this is the empirical rule for mound shaped data shown below. We will explore this rule in more detail when we study the normal distribution.

The Empirical Rule for Mound Shaped Data

The empirical rule given below applies to data sets with frequency distributions that are mound shaped and symmetric, like the one shown below.



Mound shaped distributions are very important because they frequently occur as population distributions (see Normal distribution/central limit theorem later).



If the data has a frequency distribution which is mound shaped and symmetric, we have the following empirical rule:

- Approximately 68% of the measurements will fall within 1 standard deviation of the mean i.e. within the interval $(\bar{x} - s, \bar{x} + s)$ for a sample and $(\mu - \sigma, \mu + \sigma)$ for a population.
- Approximately 95% of the measurements will fall within 2 standard deviations of the mean, i.e. within the interval $(\bar{x} - 2s, \bar{x} + 2s)$ for samples and $(\mu - 2\sigma, \mu + 2\sigma)$ for a population.
- Approximately 99.7% of the measurements(essentially all) will fall within 3 standard deviations of the mean.

Numerical Measures of Relative Standing

Quite often when interpreting a data observation, such as a baby's height and weight, we are interested in how it compares to the rest of the relevant population. Measures of relative standing describe the location of a particular measurement relative to the rest of the data. We explore some of the standard measures of relative standing below.

Z-Scores The z-score for a particular measurement in a set of data, measures how many standard deviations that measurement lies away from the mean.

Definition The **sample z-score** for a measurement x in a set of data is

$$z = \frac{x - \bar{x}}{s}$$

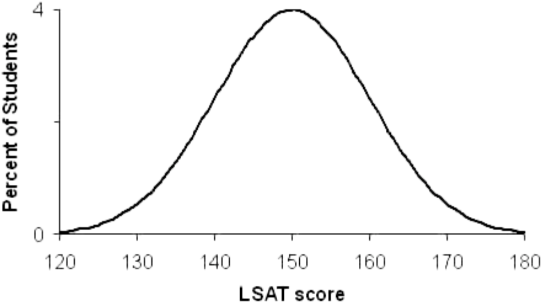
where s is the sample standard deviation.

The **population z-score** for a data measurement, x, is

$$z = \frac{x - \mu}{\sigma}$$

where σ is the population standard deviation.

Example The scores on the LSAT for a particular year have a mound shaped distribution, mean $\mu = 150$ standard deviation $\sigma = 10$. The distribution is shown below.



(a) Use the empirical rule to determine what percentage of prospective law students have z scores between -2 and 2.

From the formula, $x = \mu + z \cdot \sigma$ so a z score between -2 and 2 means you are in the interval $(\mu - 2\sigma, \mu + 2\sigma)$ and hence by the Empirical Rule 95% of the students taking the exam have z scores between -2 and 2.

(b) If you scored 175 on the exam, what would your z-score be?

From the formula, $z = \dfrac{x - \mu}{\sigma} = \dfrac{175 - 150}{10} = 2.5.$

Example In 2013 Mary was among the college bound seniors who took the SAT and ACT exams. Her composite score on the SAT was 2500 and her composite score on the ACT was 34. The national average for the composite score on the SAT among college bound seniors for that year was 1499 and the standard deviation was 319. The national average for the ACT among college bound seniors for that year was 22.5 and the standard deviation was 4.9.

Find Mary's z-score for both exams and use the z-scores to compare Mary's performance on both exams.

$$z_{\text{SAT}} = \frac{x-\mu}{\sigma} = \frac{2500-1499}{319} \approx 3.1379310345.$$

$$z_{\text{ACT}} = \frac{x-\mu}{\sigma} = \frac{34-22.5}{4.9} \approx 2.3469387755.$$

Mary did better on the SAT.

Rank

Rank We can also use rank to measure relative standing, by ranking the data as 1st, 2nd, 3rd, according to the size of the data measurement. This is commonly used in racing, where a lower time leads to a higher position, also in many competitions a higher number of points or wins leads to a higher rank. When two data measurements are the same (a tie) we can give both the same rank and skip a rank. A closely related measure of relative standing is given by the percentile:

Percentiles

Recall that the **median** of a set of data is number for which 50% of the measurements lie at or below the median and 50% lie at or above it. This is the 50th percentile of the distribution.

For any set of n measurements, (arranged in ascending or descending order), the pth percentile is a number such that p% of the measurements fall at or below that number and $(100 - p)$% of the measurements fall above it. The calculation of percentiles is not well defined and there are a few conventions which one might adopt for choosing a value for a percentile.

We will adopt a relatively simple convention which mostly agrees with our calculation of the median from before. (If the sample size is odd, it agrees precisely and if the sample size is even it is close.) For a set of data of size N, to calculate the $P-$th percentile we order our data from smallest to largest and choose the $n-$th data point on the list, where n is the nearest integer above (or equal to) $\frac{P}{100} \times N$.

Example 1 Find the 10th percentile, the 25th percentile, the 50th percentile, the 75th percentile and the 90th percentile of the following set of 20 exam scores:

$$60, \quad 71, \quad 85, \quad 99, \quad 100, \quad 76, \quad 98, \quad 61, \quad 75, \quad 82, \quad 95,$$

$$72, \quad 88, \quad 61, \quad 72, \quad 80, \quad 100, \quad 90, \quad 60, \quad 70.$$

In this example there are 20 data points and in increasing order they are

$$60, 60, 61, 61, 70, 71, 72, 72, 75, 76, 80,$$

$$82, 85, 88, 90, 95, 98, 99, 100, 100$$

For the 10th percentile, $\frac{P}{100} \cdot N = \frac{10}{100} \cdot 20 = 2.$ Hence 60 is the answer.

For the 25th percentile, $\frac{P}{100} \cdot N = \frac{25}{100} \cdot 20 = 5.$ Hence 70 is the answer.

For the 50th percentile, $\frac{P}{100} \cdot N = \frac{50}{100} \cdot 20 = 10.$ Hence 76 is the answer. Notice the 50th percentile is not quite the median, which in this case is $\frac{76+80}{2} = 78.$

$$60, 60, 61, 61, 70, 71, 72, 72, 75, 76, 80,$$
$$82, 85, 88, 90, 95, 98, 99, 100, 100$$

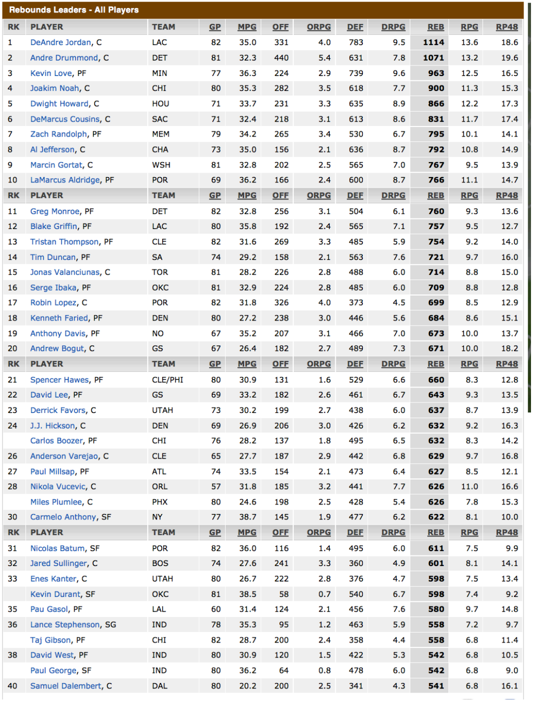
For the 75th percentile, $\frac{P}{100} \cdot N = \frac{75}{100} \cdot 20 = 15$. Hence 90 is the answer. For the 90th percentile, $\frac{P}{100} \cdot N = \frac{90}{100} \cdot 20 = 18$. Hence 100 is the answer.

Just to have an example where the relevant calculation is not an integer, for the 37th percentile, $\frac{P}{100} \cdot N = \frac{37}{100} \cdot 20 = 7.4$. Hence we want the 8th number on the list or 72.

Example On the next page you will find list of the top forty players from the NBA with the number of rebounds for the 2013-2014 regular season for each given in the highlighted column. The players are ranked 1-40 according to the number of rebounds.

What is the 95-th percentile for the number of rebounds among all NBA players for the 2013-2014 regular season?

There are 40 data points so $\frac{P}{100} \cdot N = \frac{95}{100} \cdot 40 = 38$. Hence we need to count 38 rows starting at the bottom. For large numbers like this it would be easier to count from the top. The formula is $40 - 38 + 1 = 3$ so the answer is 963. The $+1$ comes from the fact that we start counting with 1.



Rebounds Leaders - All Players

RK	PLAYER	TEAM	GP	MPG	OFF	ORPG	DEF	DRPG	REB	RPG	RP48
1	DeAndre Jordan, C	LAC	82	35.0	331	4.0	783	9.5	**1114**	13.6	18.6
2	Andre Drummond, C	DET	81	32.3	440	5.4	631	7.8	**1071**	13.2	19.6
3	Kevin Love, PF	MIN	77	36.3	224	2.9	739	9.6	**963**	12.5	16.5
4	Joakim Noah, C	CHI	80	35.3	282	3.5	618	7.7	**900**	11.3	15.3
5	Dwight Howard, C	HOU	71	33.7	231	3.3	635	8.9	**866**	12.2	17.3
6	DeMarcus Cousins, C	SAC	71	32.4	218	3.1	613	8.6	**831**	11.7	17.4
7	Zach Randolph, PF	MEM	79	34.2	265	3.4	530	6.7	**795**	10.1	14.1
8	Al Jefferson, C	CHA	73	35.0	156	2.1	636	8.7	**792**	10.8	14.9
9	Marcin Gortat, C	WSH	81	32.8	202	2.5	565	7.0	**767**	9.5	13.9
10	LaMarcus Aldridge, PF	POR	69	36.2	166	2.4	600	8.7	**766**	11.1	14.7
RK	PLAYER	TEAM	GP	MPG	OFF	ORPG	DEF	DRPG	REB	RPG	RP48
11	Greg Monroe, PF	DET	82	32.8	256	3.1	504	6.1	**760**	9.3	13.6
12	Blake Griffin, PF	LAC	80	35.8	192	2.4	565	7.1	**757**	9.5	12.7
13	Tristan Thompson, PF	CLE	82	31.6	269	3.3	485	5.9	**754**	9.2	14.0
14	Tim Duncan, PF	SA	74	29.2	158	2.1	563	7.6	**721**	9.7	16.0
15	Jonas Valanciunas, C	TOR	81	28.2	226	2.8	488	6.0	**714**	8.8	15.0
16	Serge Ibaka, PF	OKC	81	32.9	224	2.8	485	6.0	**709**	8.8	12.8
17	Robin Lopez, C	POR	82	31.8	326	4.0	373	4.5	**699**	8.5	12.9
18	Kenneth Faried, PF	DEN	80	27.2	238	3.0	446	5.6	**684**	8.6	15.1
19	Anthony Davis, PF	NO	67	35.2	207	3.1	466	7.0	**673**	10.0	13.7
20	Andrew Bogut, C	GS	67	26.4	182	2.7	489	7.3	**671**	10.0	18.2
RK	PLAYER	TEAM	GP	MPG	OFF	ORPG	DEF	DRPG	REB	RPG	RP48
21	Spencer Hawes, PF	CLE/PHI	80	30.9	131	1.6	529	6.6	**660**	8.3	12.8
22	David Lee, PF	GS	69	33.2	182	2.6	461	6.7	**643**	9.3	13.5
23	Derrick Favors, C	UTAH	73	30.2	199	2.7	438	6.0	**637**	8.7	13.9
24	J.J. Hickson, C	DEN	69	26.9	206	3.0	426	6.2	**632**	9.2	16.3
	Carlos Boozer, PF	CHI	76	28.2	137	1.8	495	6.5	**632**	8.3	14.2
26	Anderson Varejao, C	CLE	65	27.7	187	2.9	442	6.8	**629**	9.7	16.8
27	Paul Millsap, PF	ATL	74	33.5	154	2.1	473	6.4	**627**	8.5	12.1
28	Nikola Vucevic, C	ORL	57	31.8	185	3.2	441	7.7	**626**	11.0	16.6
	Miles Plumlee, C	PHX	80	24.6	198	2.5	428	5.4	**626**	7.8	15.3
30	Carmelo Anthony, SF	NY	77	38.7	145	1.9	477	6.2	**622**	8.1	10.0
RK	PLAYER	TEAM	GP	MPG	OFF	ORPG	DEF	DRPG	REB	RPG	RP48
31	Nicolas Batum, SF	POR	82	36.0	116	1.4	495	6.0	**611**	7.5	9.9
32	Jared Sullinger, C	BOS	74	27.6	241	3.3	360	4.9	**601**	8.1	14.1
33	Enes Kanter, C	UTAH	80	26.7	222	2.8	376	4.7	**598**	7.5	13.4
	Kevin Durant, SF	OKC	81	38.5	58	0.7	540	6.7	**598**	7.4	9.2
35	Pau Gasol, PF	LAL	60	31.4	124	2.1	456	7.6	**580**	9.7	14.8
36	Lance Stephenson, SG	IND	78	35.3	95	1.2	463	5.9	**558**	7.2	9.7
	Taj Gibson, PF	CHI	82	28.7	200	2.4	358	4.4	**558**	6.8	11.4
38	David West, PF	IND	80	30.9	120	1.5	422	5.3	**542**	6.8	10.5
	Paul George, SF	IND	80	36.2	64	0.8	478	6.0	**542**	6.8	9.0
40	Samuel Dalembert, C	DAL	80	20.2	200	2.5	341	4.3	**541**	6.8	16.1