MEASURES OF VARIATION

In addition to knowing what the center of our data is, we also want to know how spread out the data is. We call our techniques for this measures of variation.

Below are two sets of numbers. Both sets have a mean of 70, but clearly one data set has more variation than the other.

50, 60, 70, 75, 95

70, 70, 70, 70, 70

$$\mu = 70$$

How can we measure the variation in data sets such as the one below?

50, 60, 70, 75, 95

$$\mu = 70$$

One way we can do it is very quick, but also not very reliable in practice. It's called the range. Why is this measure of variation almost worthless?

50, 60, 70, 75, 95

$$\text{range} = \text{high - low} = 95 - 50 = 45$$

Another approach might be to find the difference between each score and the mean, and then compute the average difference. However, the table on the next slide shows us that there is a problem with this method.

50, 60, 70, 75, 95

$$\mu = 70$$

Another approach might be to find the difference between each score and the mean, and then compute the average difference.

50, 60, 70, 75, 95

$$\mu = 70$$

x	**x - μ**
50	-20
60	-10
70	0
75	5
95	25

Another approach might be to find the difference between each score and the mean, and then compute the average difference.

50, 60, 70, 75, 95

$$\mu = 70$$

x	**x - µ**
50	-20
60	-10
70	0
75	5
95	25

$$\sum (x - \mu) = 0$$

The problem is that the negative and the positive differences completely cancel each other out giving us a sum of zero, and this will happen every time.

50, 60, 70, 75, 95

$$\mu = 70$$

x	x - μ
50	-20
60	-10
70	0
75	5
95	25

$$\sum (x - \mu) = 0$$

A way around this problem is to square the difference between each score and the mean in order to eliminate negative numbers.

50, 60, 70, 75, 95

$$\mu = 70$$

x	x - µ	(x - µ)^2
50	-20	400
60	-10	100
70	0	0
75	5	25
95	25	625
	sum =	**1150**

Next, we'll find the average squared difference, and then to somewhat undo the effect of squaring, we'll take the square root of the whole thing.

50, 60, 70, 75, 95

$$\mu = 70$$

x	x - µ	(x - µ)^2
50	-20	400
60	-10	100
70	0	0
75	5	25
95	25	625
	sum =	**1150**

This particular method of measuring variation is called the standard deviation.

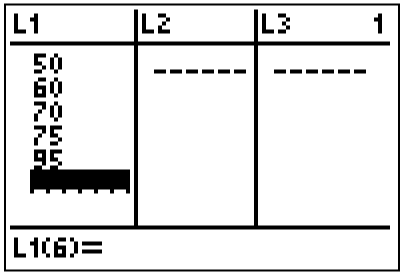
$$\text{standard deviation} = \sqrt{\frac{\sum (x-\mu)^2}{n}}$$

However, if we are finding the standard deviation of a sample, then we divide by *n-1* instead of *n*. This results in a better estimate of the population standard deviation. However, the reason why is very technical.

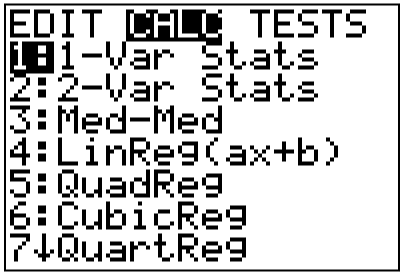
$$\text{population standard deviation} = \sigma = \sqrt{\frac{\sum (x-\mu)^2}{n}}$$

$$\text{sample standard deviation} = s = \sqrt{\frac{\sum (x-\overline{x})^2}{n-1}}$$

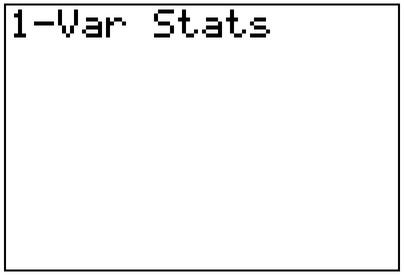
Both of these calculations are done automatically for us by our TI calculator. Enter your data into *List 1*, and go to *Stats* → *Calc.*



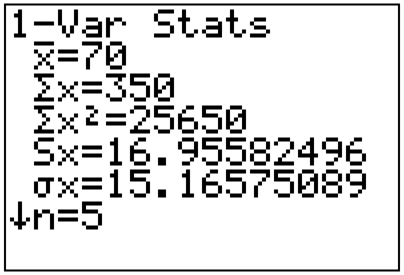
```
L1     |L2     |L3     1
  50   |-------|-------
  60   |       |
  70   |       |
  75   |       |
  95   |       |
 ------|       |
L1(6)=
```



```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```



```
1-Var Stats
```



```
1-Var Stats
 x̄=70
 Σx=350
 Σx²=25650
 Sx=16.95582496
 σx=15.16575089
↓n=5
```

$$s \approx 16.956$$

$$\sigma \approx 15.166$$

The square of the standard deviation is called the variance.

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

$$\text{sample variance} = s^2 = \frac{\sum (x - \overline{x})^2}{n - 1}$$

Now for some algebra magic.

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

$$= \frac{\sum x^2 - 2\mu \sum x + n\mu^2}{n} = \frac{\sum x^2 - 2\frac{\sum x}{n}\sum x + n\left(\frac{\sum x}{n}\right)^2}{n}$$

$$= \frac{\sum x^2 - \frac{2\left(\sum x\right)^2}{n} + \frac{\left(\sum x\right)^2}{n}}{n} = \frac{\sum x^2 - \frac{\left(\sum x\right)^2}{n}}{n}$$

Take the square root and we get what we call the raw score formula for the population standard deviation.

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

And, of course, the raw score formula for the sample standard deviation is very similar.

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

At this point, we should now see if we can determine how to calculate variance and standard deviation for a probability distribution. On the one hand, we know that variance is defined by the formula below.

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

This formula basically computes the average squared deviation from the mean. But on the other hand, if the formula in blue below gives the average value for the distribution, then the modified one in red should give the average squared deviation, i.e. the variance.

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

$$E = \sum [x \cdot P(x)] \qquad \sigma^2 = \sum \left[(x-\mu)^2 \cdot P(x) \right]$$

Now we just need to do a little algebra to get this in a better form for computations.

$$\sigma^2 = \sum\left[(x-\mu)^2 \cdot P(x)\right] = \sum\left[(x^2 - 2x\mu + u^2) \cdot P(x)\right]$$

$$= \sum\left[x^2 P(x) - 2x\mu P(x) + \mu^2 P(x)\right]$$

$$= \sum\left[x^2 P(x)\right] - \sum[2x\mu P(x)] + \sum\left[\mu^2 P(x)\right]$$

$$= \sum\left[x^2 P(x)\right] - 2\mu\sum[x \cdot P(x)] + \mu^2\sum[P(x)]$$

$$= \sum\left[x^2 P(x)\right] - 2\mu \cdot \mu + \mu^2 = \sum\left[x^2 P(x)\right] - 2\mu^2 + \mu^2$$

$$= \sum\left[x^2 \cdot P(x)\right] - \mu^2 = \sum\left[x^2 \cdot P(x)\right] - \left(\sum[x \cdot P(x)]\right)^2$$

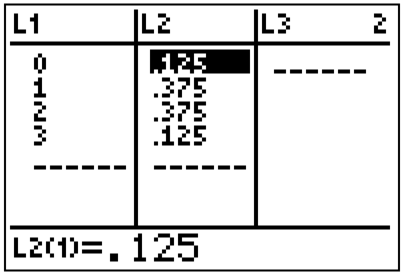
Now lets give it a try using our probability distribution for the coin flipping experiment.

x = number of heads	*P(x)*	x*P(x)	x^2	x^2*P(x)
0	0.125	0	0	0
1	0.375	0.375	1	0.375
2	0.375	0.750	4	1.5
3	0.125	0.375	9	1.125
		1.5		3

$$\mu = \sum\left[x \cdot P(x)\right] = 1.5$$

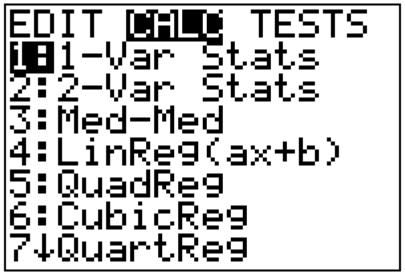
$$\sigma = \sqrt{\sum\left[x^2 \cdot P(x)\right] - \mu^2} = \sqrt{.75} \approx .8660254038$$

Of course, there is one other way to do this which now should seem very convenient.

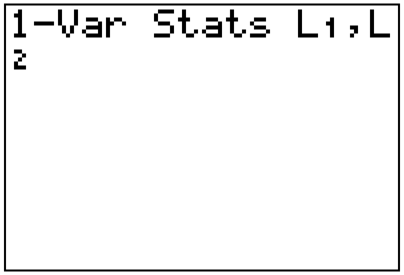


L1	L2	L3 2
0	.125	------
1	.375	
2	.375	
3	.125	
------	------	

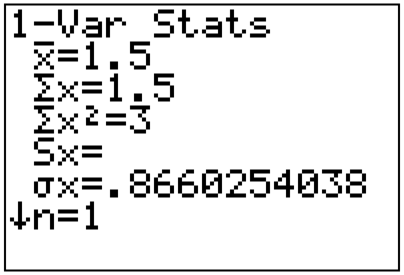
L2(1)=.125



```
EDIT CALC TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7↓QuartReg
```



```
1-Var Stats L₁,L
₂
```



```
1-Var Stats
 x̄=1.5
 Σx=1.5
 Σx²=3
 Sx=
 σx=.8660254038
↓n=1
```

$$\mu = \sum\left[x \cdot P(x)\right] = 1.5$$

$$\sigma = \sqrt{\sum\left[x^2 \cdot P(x)\right] - \mu^2} = \sqrt{.75} \approx .8660254038$$

The *standard deviation* of a binomial distribution is a simple formula, but we will give it without proof.

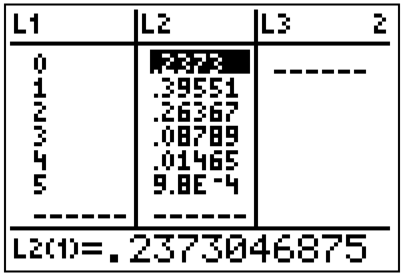
$$\mu = np$$

$$(5)(.25) = 1.25$$

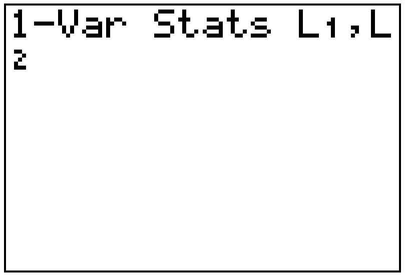
$$\sigma = \sqrt{npq}$$

$$\sqrt{5 \cdot .25 \cdot .75} \approx .96825$$

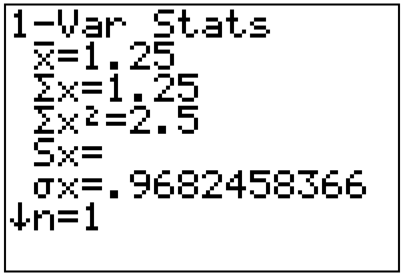
And of course, this can also be done just using the calculator.



```
L1       L2       L3      2
  0      .2373    ------
  1      .39551
  2      .26367
  3      .08789
  4      .01465
  5      9.8E-4
 ------  ------
L2(1)=.2373046875
```



```
1-Var Stats L₁,L
₂
```



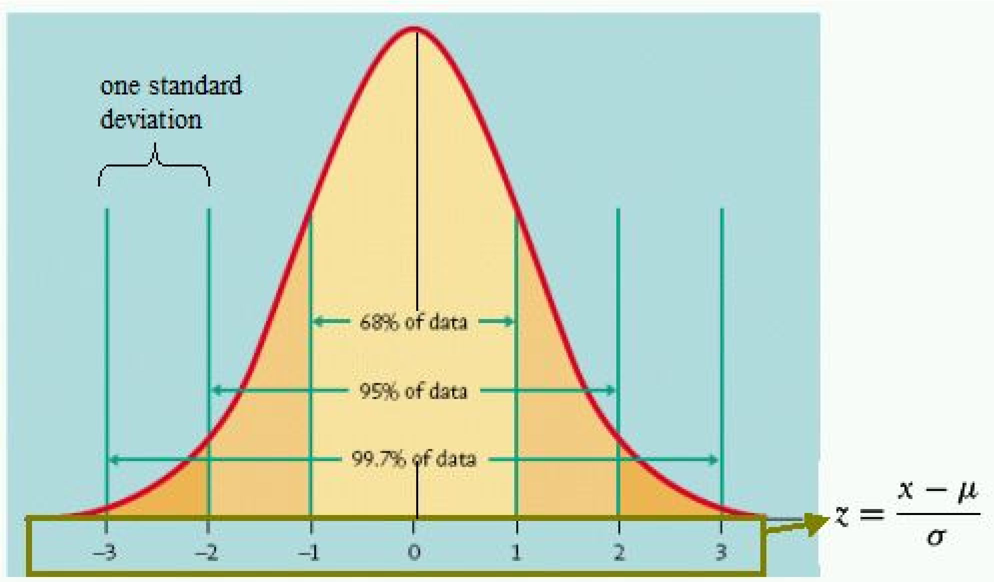
```
1-Var Stats
 x̄=1.25
 Σx=1.25
 Σx²=2.5
 Sx=
 σx=.9682458366
↓n=1
```

An interesting result is Chebyshev’s Theorem that says that for any type of distribution of data, the proportion of that data that lies within *k* standard deviations of the mean, for *k>1*, is at least $1 - 1/k^2$.

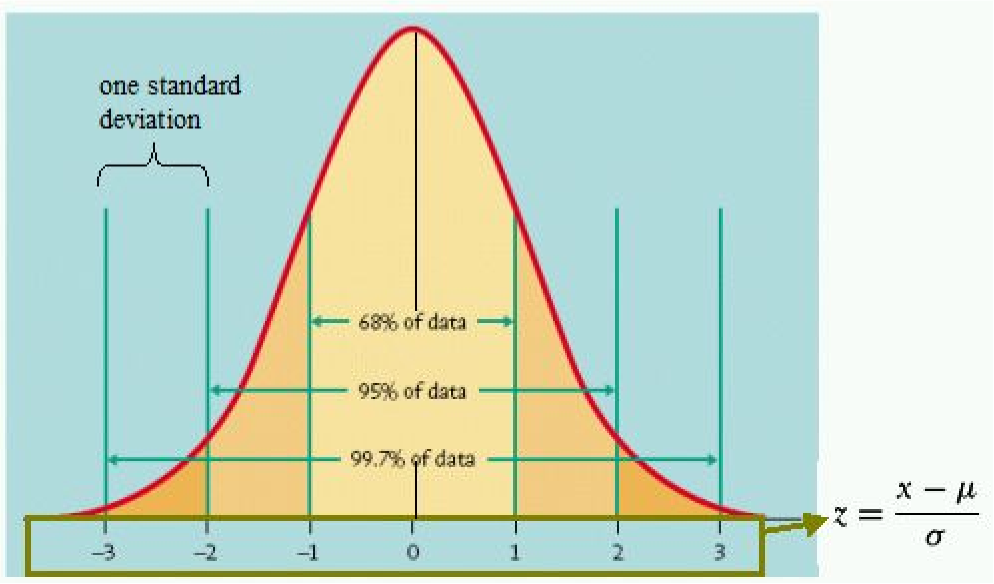
For $k = 2$, $1 - \frac{1}{2^2} = \frac{3}{4} = 75\%$ of the data (at least) lies within 2 standard deviations of the mean.

For $k = 3$, $1 - \frac{1}{3^2} = \frac{8}{9} \approx 89\%$ of the data (at least) lies within 3 standard deviations of the mean.

Often, however, we can do better than Chebyshev’s Theorem because frequently data has a bell-shaped distribution known as the normal curve.



In a normal distribution, 68% of the data is within 1 standard deviation of the mean, 95% is within 2 standard deviations, and 99.7% is within 3 standard deviations of the mean.


one standard deviation
68% of data
95% of data
99.7% of data
–3
–2
–1
0
1
2
3
$z = \frac{x - \mu}{\sigma}$