Normal Models

Donna Dietz

American University

dietz@american.edu

STAT 202 - Spring 2020

In Statistics, no model gets more attention than the Normal Model.

This is partly historical, partly pedagogical, and partly because it's actually quite useful!

We will keep learning more about it as we progress through the semester.

Normal Models

- σ is the population standard deviation
- s is the sample standard deviation
- s is actually an estimator for σ
- s is not the standard deviation of the sample itself

Parameters (Population)

- σ standard deviation
- μ mean
- p proportion

Statistics (Sample)

- s standard deviation
- \bar{x} mean
- *p̂* proportion

- Find the mean of the data.
- Write data in a column.
- Next column: subtract mean from data
- Next column: square previous column
- Add all squared values.
- Divide this sum by *n*, number of data elements. (This is σ^2 .)
- Note that σ^2 is called the variance.
- Take the square root of the variance to get the standard deviation.

1	
2	
3	
4	
5	

1	-2	
2	-1	
3	0	
4	1	
5	2	

1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

Add: 4+1+0+1+4 = 10

Add: 4+1+0+1+4 = 10Divide: 10/5 = 2 (This is σ^2 , the variance.) Add: 4+1+0+1+4 = 10Divide: 10/5 = 2 (This is σ^2 , the variance.) Square root: $\sigma = \sqrt{2} \approx 1.4142$ Finding s is almost the same as finding σ , except that we divide by n-1.

- Find the mean of the data.
- Write data in a column.
- Next column: subtract mean from data
- Next column: square previous column
- Add all squared values.
- Divide this sum by n 1. (This is s^2 .)
- You guessed it! s^2 is called the variance.
- Take the square root of the variance to get the standard deviation.

1	
2	
3	
4	
5	

1	-2	
2	-1	
3	0	
4	1	
5	2	

1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

1	-2	4
2	-1	1
3	0	0
4	1	1
5	2	4

Add: 4+1+0+1+4 = 10

Add: 4+1+0+1+4 = 10Divide: 10/4 = 2.5 (This is s^2 , the variance.) Add: 4+1+0+1+4 = 10Divide: 10/4 = 2.5 (This is s^2 , the variance.) Square root: $s = \sqrt{2.5} \approx 1.5811$ The value of s will always be larger than the value of σ calculated on thes same data.

This is because s is presuming you have a sample and you want to guess what σ is in the whole population.

However, when you calculate σ , you are only doing on an entire population!

Sometimes, even if your whole population seems to be in front of you, you will imagine it as a sample of a larger population.

Let's pretend I am doing an experiment. I collect 25 little frogs from the wild and bring them into my house in a fishtank and feed them, give them water, etc. I want to know if they will get bigger or stay the same size.

Was that a sample or a population?

You could argue that I have the entire set of frogs that I care about, or you could argue that really, I took a sample of frogs, and I am using them to infer what would happen to any of the frogs out there.

If I took the standard deviation of a classroom's worth of exams, I could make the argument that the class is just a sample of all the students who might have taken this course with me but ended up in other sections. So, then it's sample, even though I have all the data I really care about.

If I were to add 5 to every value in $\{1,2,3,4,5\}$ and get

 $\{6,7,8,9,10\}$

what would happen to the mean (μ or \bar{x})?

What would happen to σ and s?

You shouldn't bother to recalcualte all that, right? The mean will shift by the constant you added. The values of s and σ will stay the same. Why? If I were to multiply every value in $\{1,2,3,4,5\}$ by 10 and get

 $\{10, 20, 30, 40, 50\}$

what would happen to the mean (μ or \bar{x})?

What would happen to σ and s?

The mean will multiply by the 10. Why?

The values of s and σ will also multiply by 10. Why?

10	-20	400
20	-10	100
30	0	0
40	10	100
50	20	400

Add: 400+100+0+100+400 = 1000This sum is 100 times larger than before!

Divide by 5, take square root: $\sqrt{\frac{1000}{5}}=\sqrt{200}\approx 14.142$

Taking the square root brings the difference to 10.

We use formulas to make things easier to remember, because all that discussion is really hard to rehearse each time. Once we can agree a formula is correct, it doesn't take much space, and it's easier to refer to it!

$$\mu_{(x+c)} = \mu_x + c$$
$$\mu_{(Kx)} = K \times \mu_x$$

$$\sigma_{(x+c)} = \sigma_x$$
$$\sigma_{(Kx)} = K \times \sigma_x$$

We use the Normal Distribution so much that we abbreviate it this way:

$N(\mu, \sigma)$

Where μ is the mean and σ is the standard deviation.

If you shift your data using a line, what happens to the mean and standard deviation?

- The mean shifts according to the function which is a line
- The standard deviation is multiplied by the slope only

If you have temperature data in Celcius N(50, 10), what is the distribution in Fahrenheit if $F = \frac{9}{5}C + 32$? If your Celcius data look like N(50, 10), your Fahrenheit data will have a distribution of N(122, 18).

$$\mu_{(A+B)} = \mu_A + \mu_B$$

Means do what you might expect them to do. We use them in everyday life, so we are familiar with them!

Let's say you plan to shop at Target, and expect to spend N(100, 12). Then, you plan to stop at a grocery store and spend N(50, 5). How much do you expect to spend overall?

Of course, you expect to spend \$150 on average, right?

But what if you want the standard deviation?

Answer: Then you're thinking about the wrong thing!

Variances add!!! STANDARD DEVIATIONS DO NOT ADD!!!

$$\sigma_{(A+B)}^2 = \sigma_A^2 + \sigma_B^2$$

We know $\sigma_{Target} = 12$ and $\sigma_{grocery} = 5$.

Square them both to get the variances, then add them! Then, square root to get back to the standard deviation you wanted in the first place.

$$\sigma^{2}_{Target} = 144$$

$$\sigma^{2}_{grocery} = 25$$

$$\sigma^{2}_{Target+grocery} = 169$$

$$\sigma_{Target+grocery} = 13$$

It is easier, concpetually, to remind yourself that variances are what adds, but sometimes you just want to see a formula, even if the formula makes things harder to understand.

$$\sigma_{(A+B)} = \sqrt{\sigma_A^2 + \sigma_B^2}$$

What if you are visiting three stores?

You still go back to variances, add them as you wish, then take the square root of all of that!

You are going to a movie at a mall. (Yes, malls still exist!) You think parking will cost about N(10,3), the movies will cost about N(25,4) and you plan to pick up something at the mall so maybe N(120,12). Assume the variables (purchases) are each independent and don't interact in any way.

Determine your overall expected expenses, then figure out the standard deviation for them.

The mean is 10+25+120 = 155 as you would expect. The variance is $3^2 + 4^2 + 12^2$ or 9 + 16 + 144 which is 169. So the standard deviation is the square root of this, or 13. If you are unconvinced that this is the way to deal with adding up variation, that's ok. You would need to do out several examples by hand before it would make intuitive sense, but let me try to explain.

If you try to simply add the standard deviations together, you would not get the benefit of the fact that two variables are just as likely to wind up on the same sides of their means as they are to wind up on opposite sides of them. This is all held together by **independence** of variables. If they are connected somehow, none of this works, and we need other formulas!

You have a pile of apples and you are putting 9 at a time into some bags. The distribution of the apples' weights are N(6,2) ounces. What will the distribution of the bags' weights be?

You know that 9 apples at 6 ounces would be 54 ounces typically. (This is 3 pounds 6 ounces if you prefer it that way.)

For the standard deviation, you have to think in terms of variances first!

$$\sigma^{2} = 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 + 4 = 36$$
$$\sigma = \sqrt{36} = 6$$

Notice that you can shortcut this process: Multiply the original σ by the square root of of 9, since you are bagging 9 apples. $2 \times \sqrt{9} = 2 \times 3 = 6$.

$$\sigma_{(A+B)} = \sqrt{\sigma_A^2 + \sigma_B^2}$$
$$\sigma_{(A-B)} = \sqrt{\sigma_A^2 + \sigma_B^2}$$

Even if you are subtracting, if you have independence, you still add the variances! (Don't worry, means do subtract, just like you expect!)

This is because the *noise* or *variation* is just as likely to go high or go low. This is true whether you are adding or subtracting two things. Since all normal models are really the same except for adding and multiplying, we often *normalize* our data so we can refer to them with common language. We use *z*-scores to do this.

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

Alan received a 1025 on his SAT which had a distribution of N(1000, 500) that year. Betty received a 21 on her ACT which had a distribution of N(20, 5) that year. Who did better? Who had the higher z-score?

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

Find the z-score for each student.

Alan's z-score is 0.05, while Betty's is 0.2, so Betty did better.

To reverse this process, use

$$x = \mu + z\sigma.$$

Marco received a z = 1.2 on his test which had a distribution of N(80, 10). Find his raw score.

Marco's z-score is

 $80 + 1.2 \times 10 = 92.$

MEMORY QUESTIONS Just 21 today!



















































































