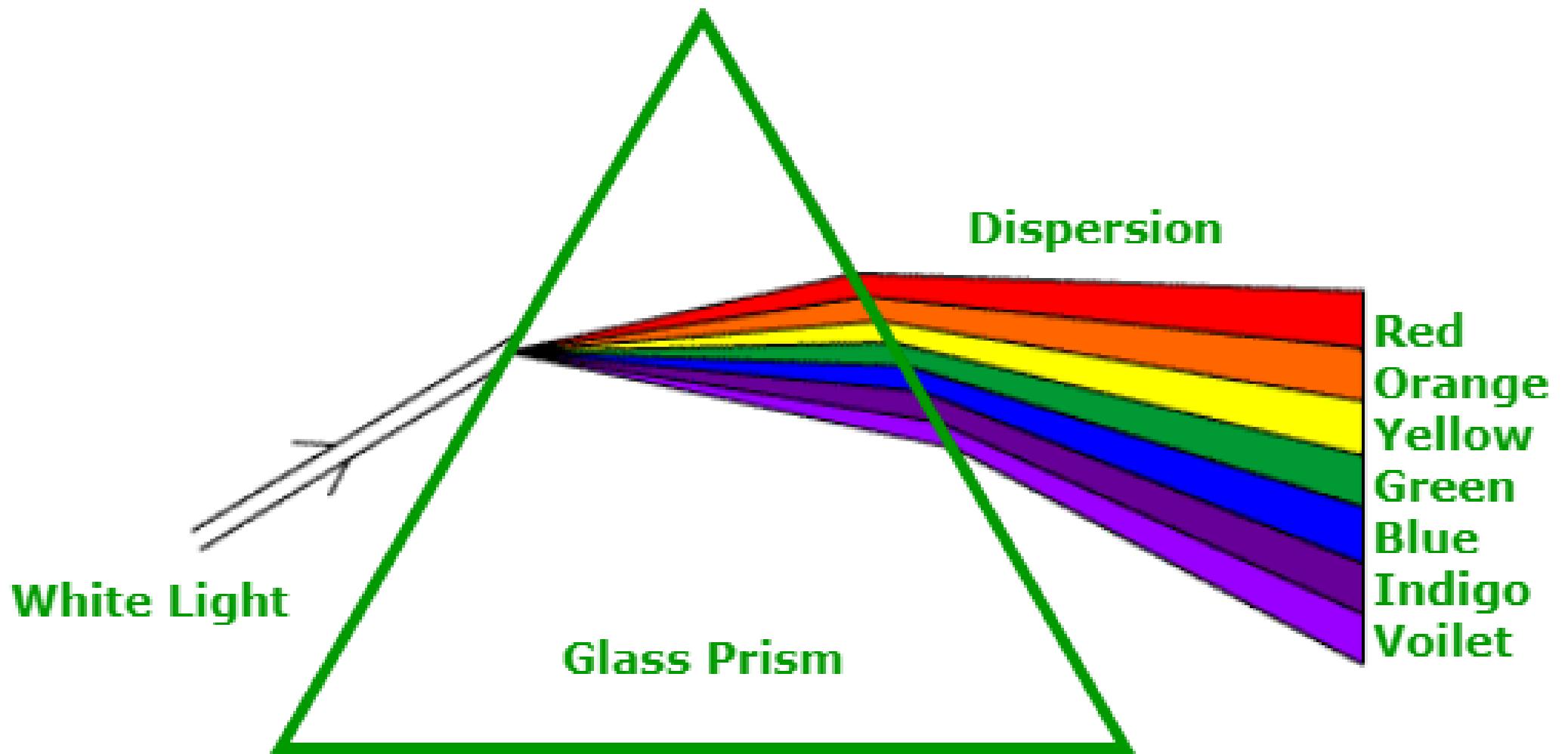


CHAPTER 5

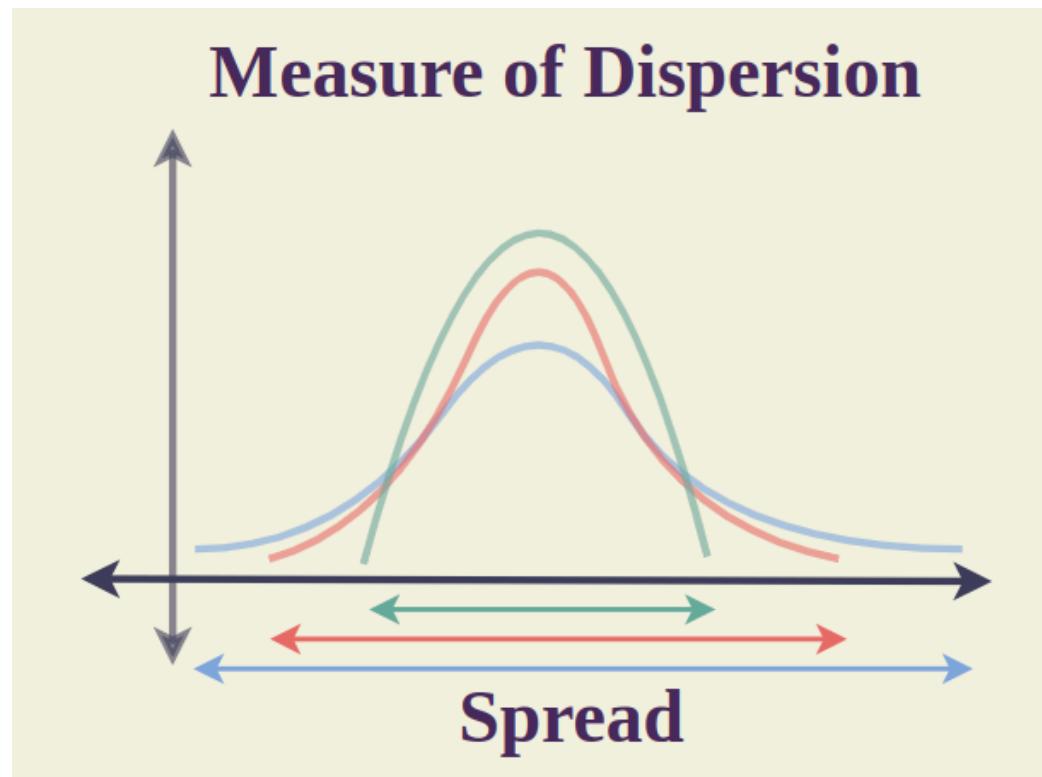
MEASURES OF DISPERSION

WHAT IS DISPERSION AS A CONCEPT?



WHAT IS DISPERSION WITHIN STATISTICS?

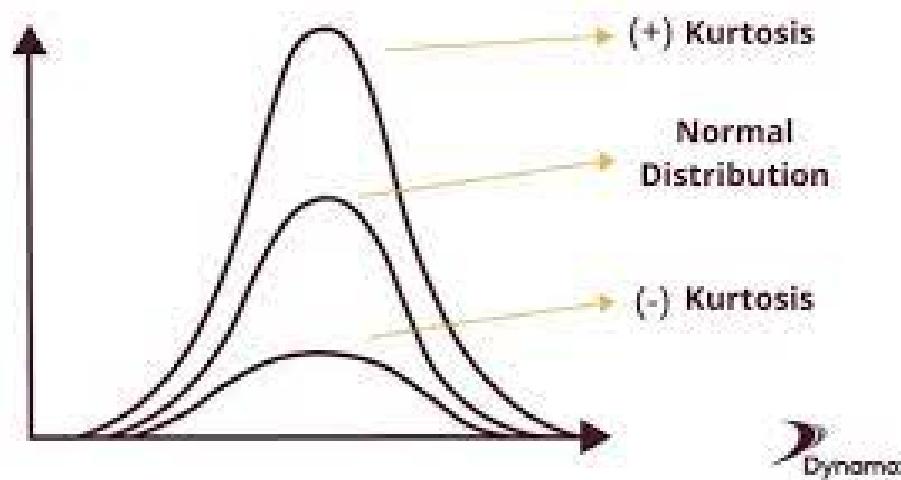
It is the amount of spread or variability among raw scores in a distribution.



KURTOSIS - THE COUSIN OF SKEWNESS

The 'sharpness' of the peak of a frequency-distribution curve.

The 'sharpness' of the peak occurs due to the tailedness (i.e., how often outliers occur) with the distribution.



TYPES OF KURTOSIS

Leptokurtosis

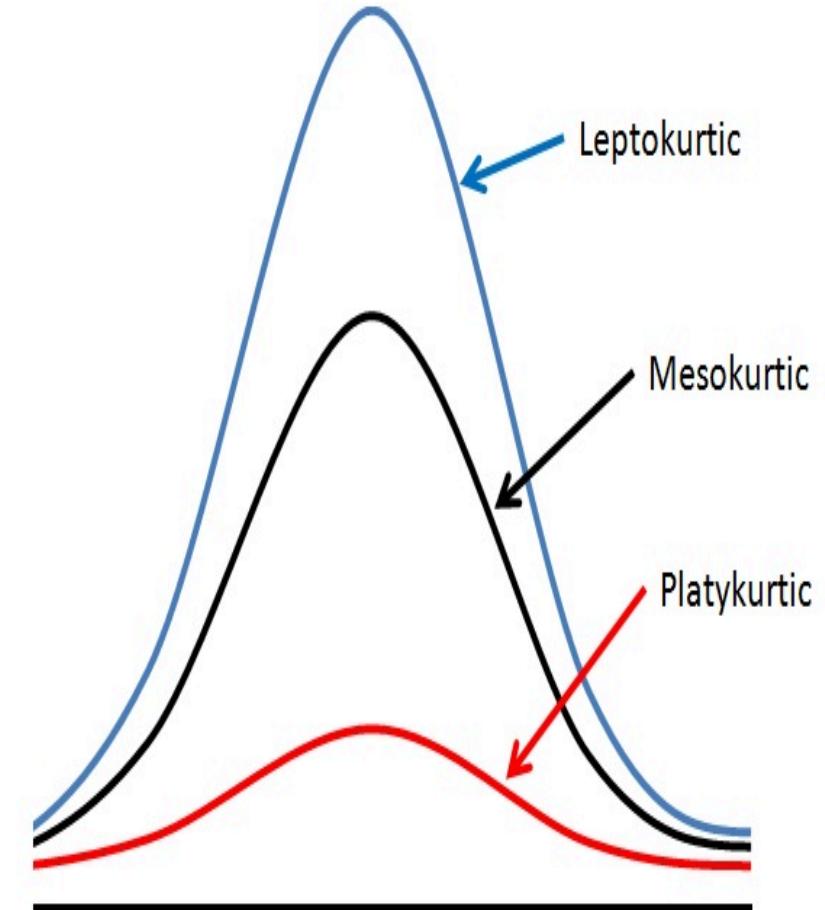
DATA CLUSTERED AROUND THE MEAN AND FEWER DATA POINTS IN THE TAILS COMPARED TO A NORMAL DISTRIBUTION

Mesokurtosis

DATA CLOSELY RESEMBLES THAT OF A NORMAL DISTRIBUTION

Platykurtosis

DATA POINTS ARE MORE SPREAD OUT, WITH FEWER VALUES CLUSTERED AROUND THE MEAN AND MORE VALUES IN THE TAILS



MEASURES OF DISPERSION FOR CATEGORICAL VARIABLES

The Variation Ratio

- A measure of dispersion.
- The proportion of cases which are not in the mode category.
- The only measure of dispersion that can be used with categorical variables

THE VARIATION RATIO

$$v = 1 - \frac{fm}{n}$$

fm = the frequency (number of cases) of the mode

n = sample size

VARIATION RATIO EXAMPLE

We asked 1,984 individuals at the University what their favorite color was. We were left with four colors: red, orange, green, and blue. What is the variation score?

colors	
Red	43
Orange	211
Green	341
Blue	1389

$$v = 1 - \frac{fm}{n}$$

RECAP

Dispersion

Kurtosis

- Leptokurtosis
- Mesokurtosis
- Platykurtosis

The Variation Ratio

MEASURES OF DISPERSION FOR CONTINUOUS VARIABLES

Range

Variance

Standard Deviation

RANGE

A measure of the span of data.

A high range value indicates there is more dispersion, and the lower the range, the less the dispersion

Range = The Maximum Value – The Minimum Value

RANGE EXAMPLE

Below is a data frame that contains the average salary within each district. Find the range.

	Salary
District1	87000
District2	91000
District3	66500
District4	98500
District5	96500
District6	97550
District7	97900
District8	97990

?

VARIANCE

A measure of how spread the observed values are from the mean.

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

s^2 = sample variance

x_i = raw value

\bar{x} = mean values of all raw scores

n = number of observations

BREAK DOWN OF EQUATION

1. Find the \bar{x}

2. Subtract the mean from each score $(x_i - \bar{x})$

3. Square the deviation score

$$(x_i - \bar{x})^2$$

4. Add the squared deviations

$$\sum(x_i - \bar{x})^2$$

5. Divide by $n - 1$

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

WHY $n-1$?

Suppose we draw n independent observations from a population (N), with an unknown population mean (μ) and unknown variance (s^2).

Ideally we would use (s^2) to find the average squared distance from the true mean,

$$s^2 = \frac{\sum(x_i - \mu)^2}{n}$$

Although we can't! 😢

Because we don't know our (μ).

WE SAMPLE IT!

Since we don't know (μ) , we use our best estimate of it which is the sample mean \bar{x} .

So let's pull out the population variable (μ) and plug in our sample variables (\bar{x}) .

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

Although another small problems pops up!

THE PROBLEM (POPULATION)

Population and the true μ

(-5, 0) * - - * - * - * - - - (0, 0) - - | - - * - - * - - * - - -

THE PROBLEM (SAMPLE)

Sample and the \bar{x}

$$(-5, 0) * - - * - * - - - - - (0, 0) - - \overset{\mu}{|} - - - - - - - - - (5, 0)$$

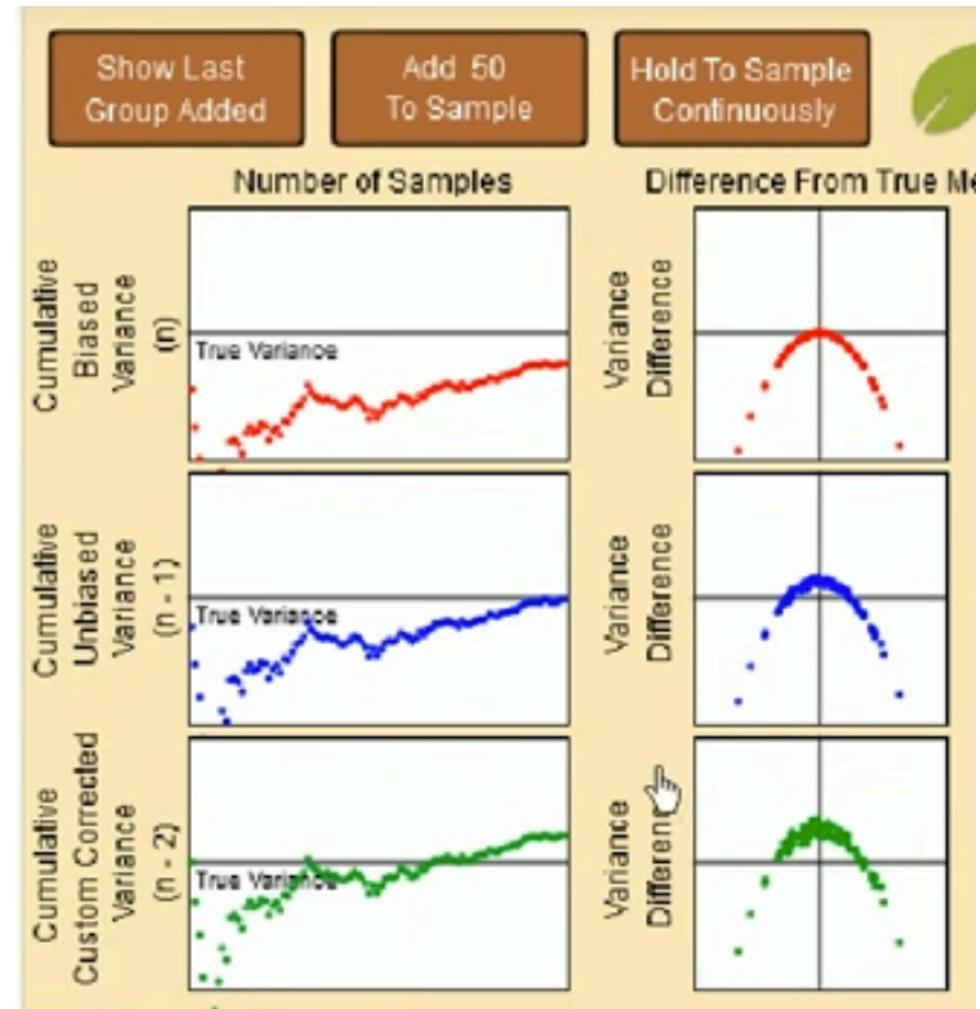
The sample mean can tend to underestimate or even overestimate the true μ

THE SOLUTION

We modify $s^2 = \frac{\sum(x_i - \bar{x})^2}{n}$ to

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

And this provides an unbiased estimator of the population variance when using a sample.



IN SUM

When we are measuring a population, divide by n . (hint: There will be a μ in the deviation score.)

When we are measuring a sample, divide by $n-1$. (hint: There will be a \bar{x} in the deviation score.)

VARIANCE EXAMPLE 1

1) FIND THE \bar{x}

2) SUBTRACT THE MEAN FROM EACH SCORE $(x_i - \bar{x})$

3) SQUARE THE DEVIATION SCORE

$$(x_i - \bar{x})^2$$

4) ADD THE SQUARED DEVIATIONS

$$\sum(x_i - \bar{x})^2$$

5) DIVIDE BY $n - 1$

	District	Salary
1	District1	832
2	District2	931
3	District3	1468
4	District4	1021
5	District5	1039
6	District6	1515
7	District7	1138
8	District8	620

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

VARIANCE EXAMPLE 2

1) FIND THE \bar{x}

2) SUBTRACT THE MEAN FROM EACH SCORE $(x_i - \bar{x})$

3) SQUARE THE DEVIATION SCORE

$$(x_i - \bar{x})^2$$

4) ADD THE SQUARED DEVIATIONS

$$\sum(x_i - \bar{x})^2$$

5) DIVIDE BY $n - 1$

	District	Salary
1	District1	983
2	District2	993
3	District3	1047
4	District4	1002
5	District5	1004
6	District6	1051
7	District7	1014
8	District8	962

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

STANDARD DEVIATION

A measure of how far observed values are from the mean.

It's simply the square root of our variance!

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

σ = standard deviation

x_i = raw value

\bar{x} = mean values of all raw scores

n = number of observations

WHY SQUARE ROOT IT?

In our variance equation, we squared the sum of the deviation score ($x_i - \bar{x}$) to get rid of our negative values.

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

But by doing that we created an output that is nonsensical to our data and thus our interpretation.

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

So by squaring it, we normalize the sum of the deviation score, and thus the values make sense.

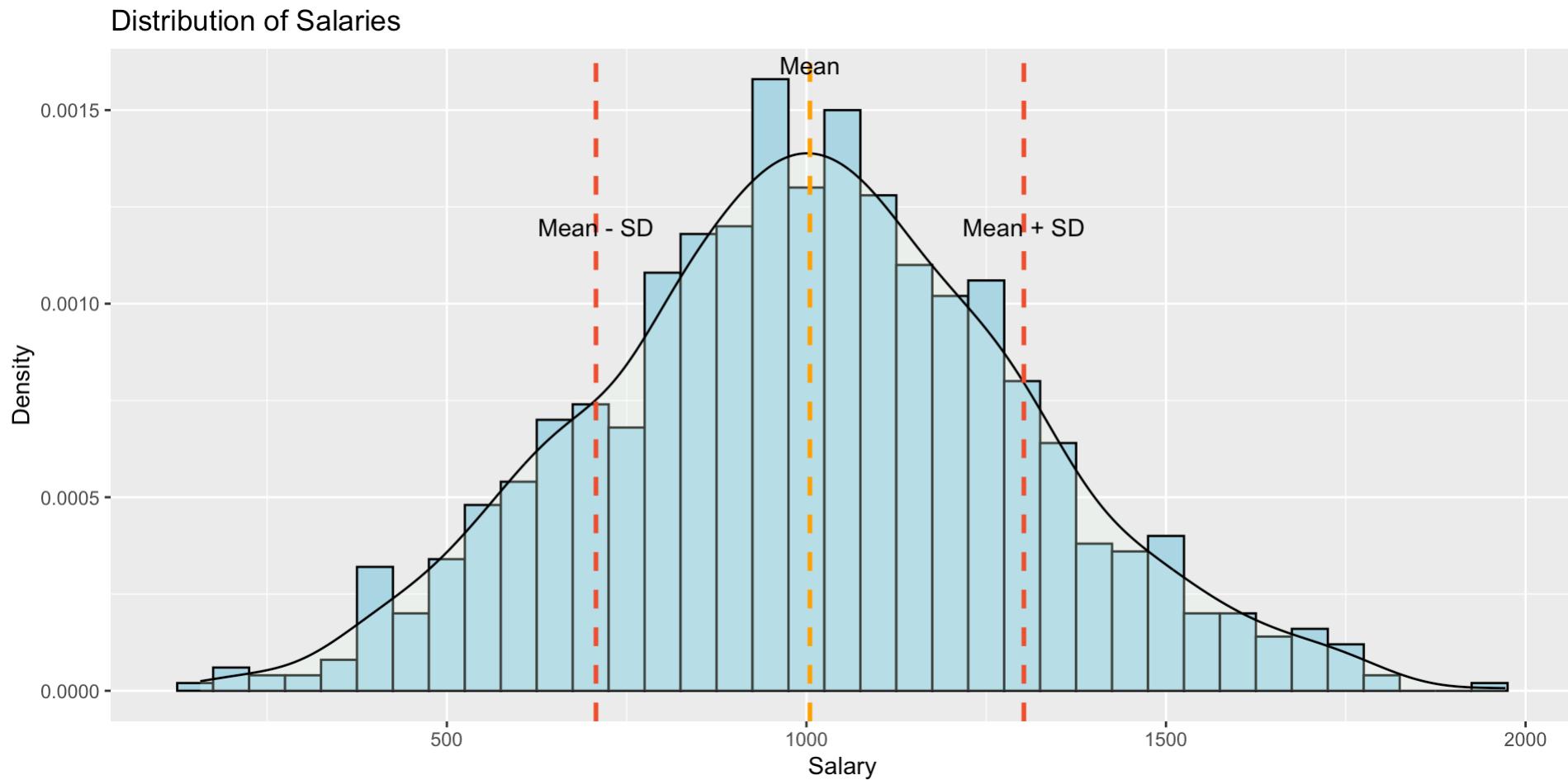
STANDARD DEVIATION EXAMPLE

Take our calculation from variance example 1 and square it.

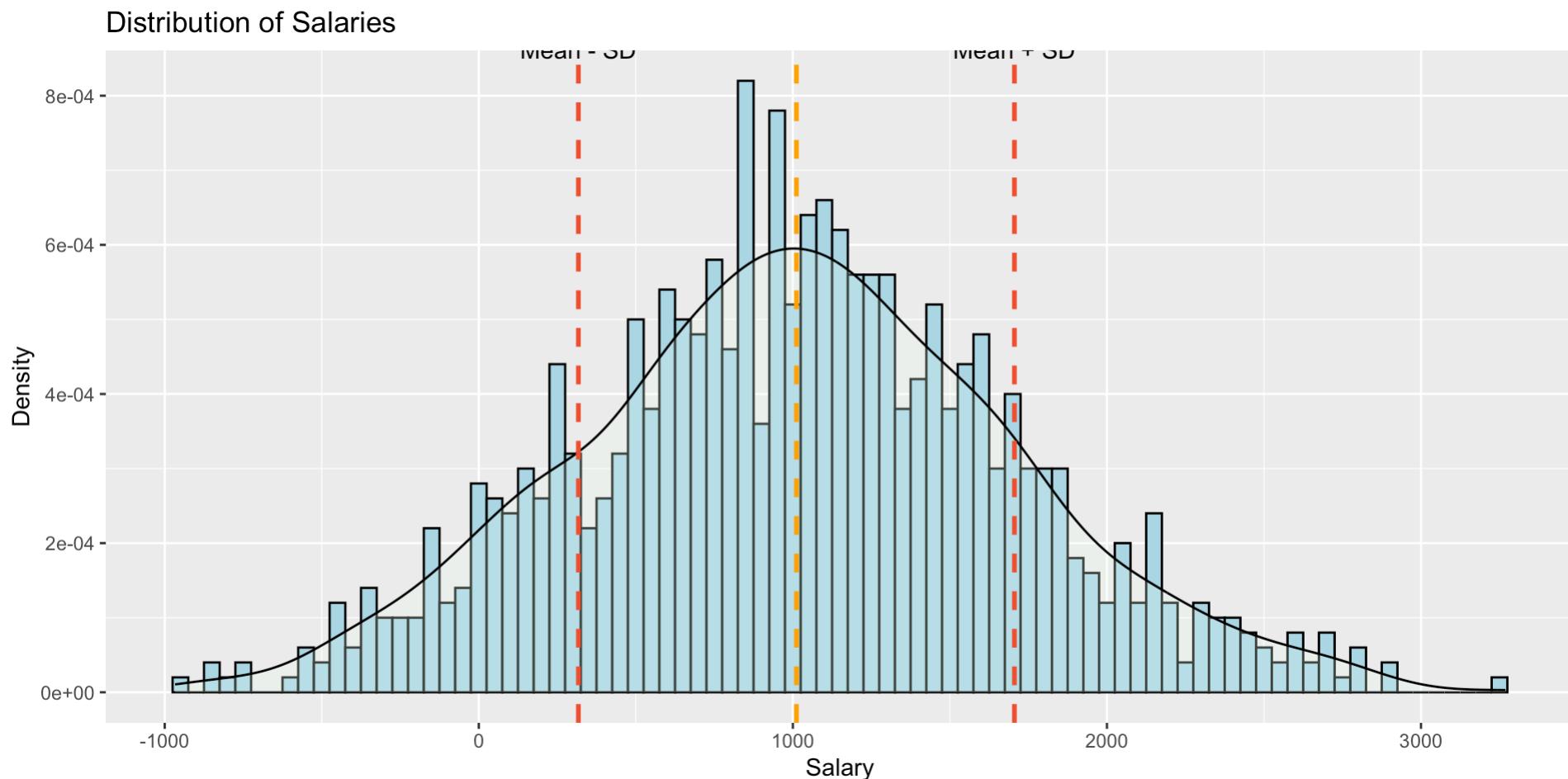
$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n - 1}}$$

District	Salary
1 District1	832
2 District2	931
3 District3	1468
4 District4	1021
5 District5	1039
6 District6	1515
7 District7	1138
8 District8	620

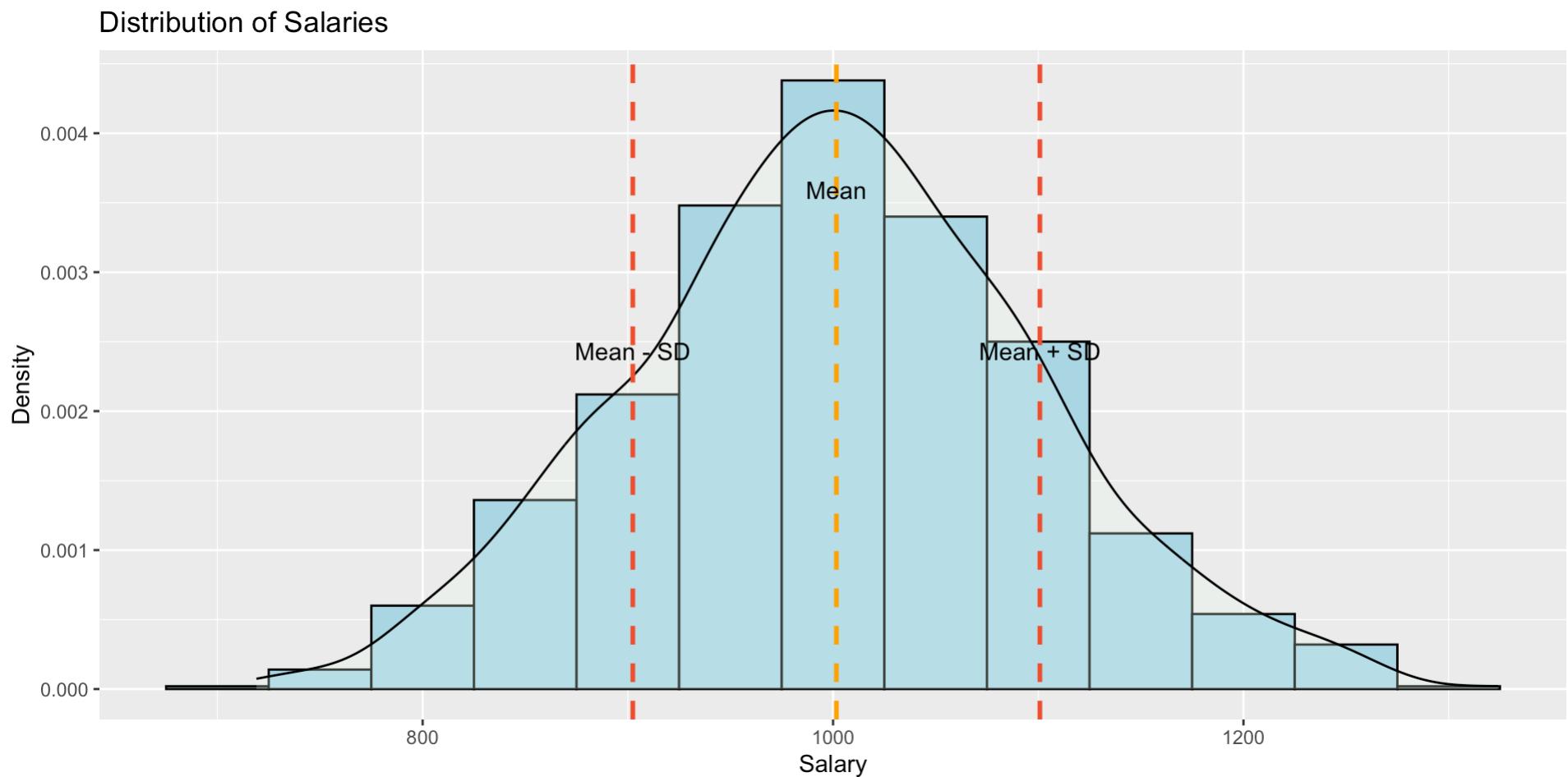
VIZUALIZE A STANDARD DEVIATION



VIZUALIZE A STANDARD DEVIATION (LARGE)

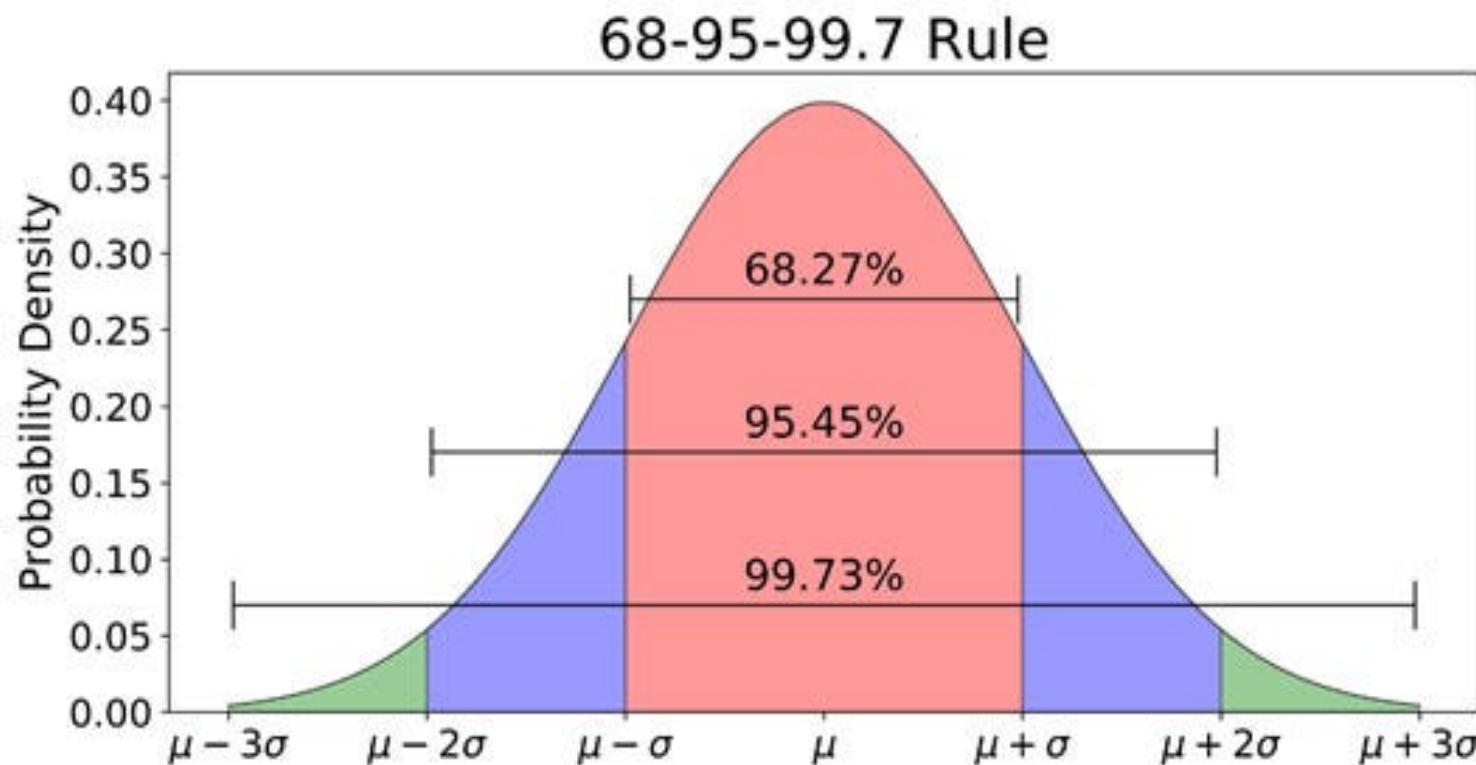


VIZUALIZE A STANDARD DEVIATION (SMALL)



THE NORMAL CURVE

On average, 68% of the sample will fall within 1 standard deviation of the mean, 95% at 2 standard deviations, and 99.7% will fall within 3 standard deviations.



HAVE WONDERFUL DAY!



TikTok

