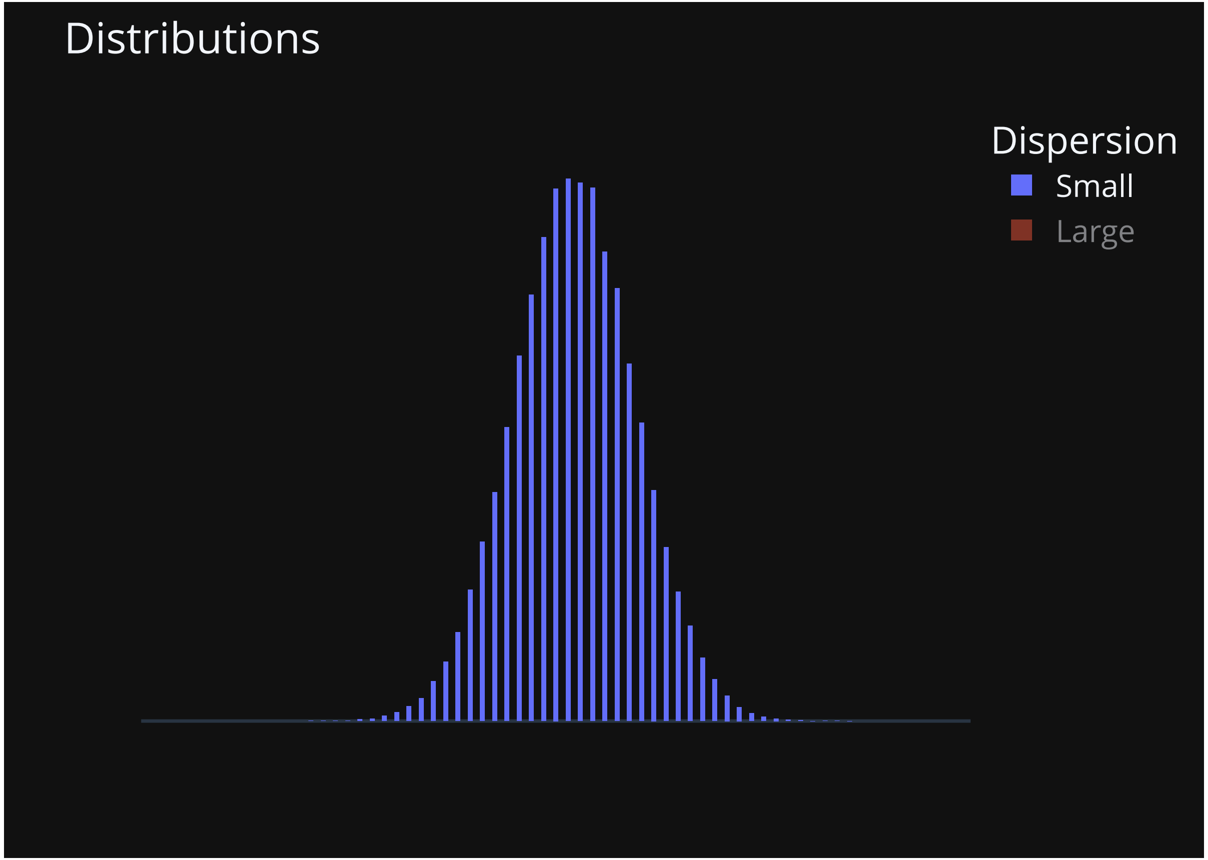
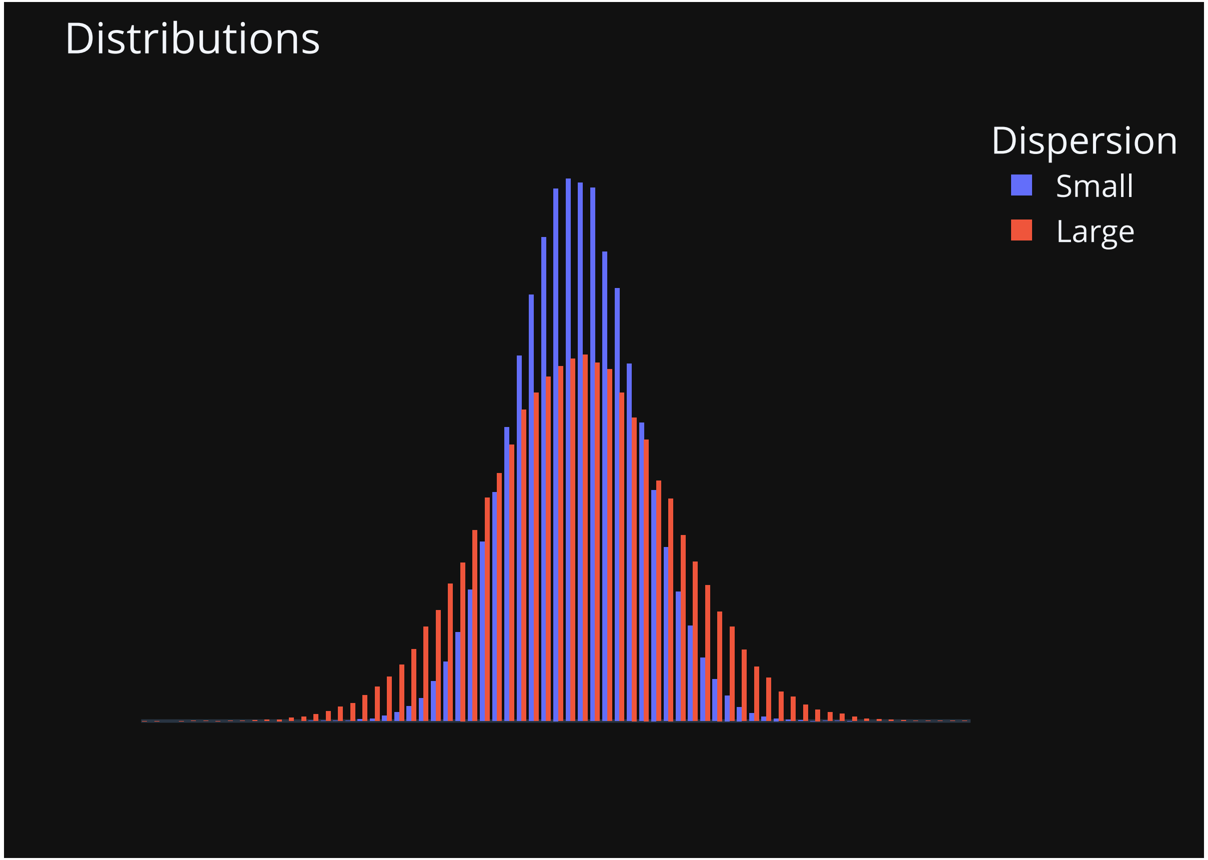
Applied Data Analytics

Statistics — Dispersion & concentration

Variance and standard deviation

Hans-Martin von Gaudecker and Aapo Stenhammar


Distributions
Dispersion
Small
Large


Distributions
Dispersion
Small
Large

Describe a DataFrame

```
df.describe().round(2)
```

Dispersion	count	mean	std	min	25%	50%	75%	max
Small	100000	0	1	-4.27	-0.67	-0	0.67	4.43
Large	100000	-0.01	1.5	-7.07	-1.01	-0	1	6.39

Variance and standard deviation

- The variance of a sample is the average squared deviation from the sample mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean

- The standard deviation is the square root of the variance

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

Degrees of freedom

	A	B
	2	1
	4	3
	6	8
Mean	?	?

	A	B
	2	1
	4	3
	?	?
Mean	4	4

Sum of squared devs., Var., and Std. Dev.

A	(A - 4)²	B	(B - 4)²
2	4	1	9
4	0	3	1
6	4	8	16
SSD	8	SSD	26
Variance	4	Variance	13
Std. Dev.	2	Std. Dev.	3.6

Pandas reductions

```
df.var()
```

```
df.std()
```

Some properties of the variance

- Linear transformations: if $y_i = a + b \cdot x_i$, then

$$s_y^2 = b^2 s_x^2$$

- Can be calculated as the difference between the average sum of squares and the squared mean:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 = \frac{n}{n-1} \left(\overline{x^2} - \overline{x}^2 \right)$$