Two Fundamentally different types of data:

Continuous

Discrete

Descriptive Statistics - continuos variables

Central tendency
MEAN (average)

Average or mean : central tendency

$$X = \begin{bmatrix} 7 \\ 10 \\ 12 \\ 12 \\ 10 \\ 5 \\ 10 \end{bmatrix}$$

$\bar{x}$: **(7 + 10 + 12 + 12 + 10 + 5 + 10) / 7 = aprox. 9.43**

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{(7 + 10 + \ldots + 5 + 10)}{N}$$

$$\frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{7}(66)$$

$$X = \begin{bmatrix} 7 \\ 10 \\ 12 \\ 12 \\ 10 \\ 5 \\ 10 \end{bmatrix}$$

Average or mean :

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{1}{7}7 + \frac{1}{7}10 + \frac{1}{7}12 + \frac{1}{7}12 + \frac{1}{7}10 + \frac{1}{7}5 + \frac{1}{7}10$$

$$= \frac{1}{7}7 + \frac{3}{7}10 + \frac{2}{7}12 + \frac{1}{7}5$$

$$= (0.14)7 + (0.43)10 + (0.29)12 + (0.14)5$$

$$\bar{x} = \sum_{\textbf{All x}} x = P(x) * x$$

Weighted average

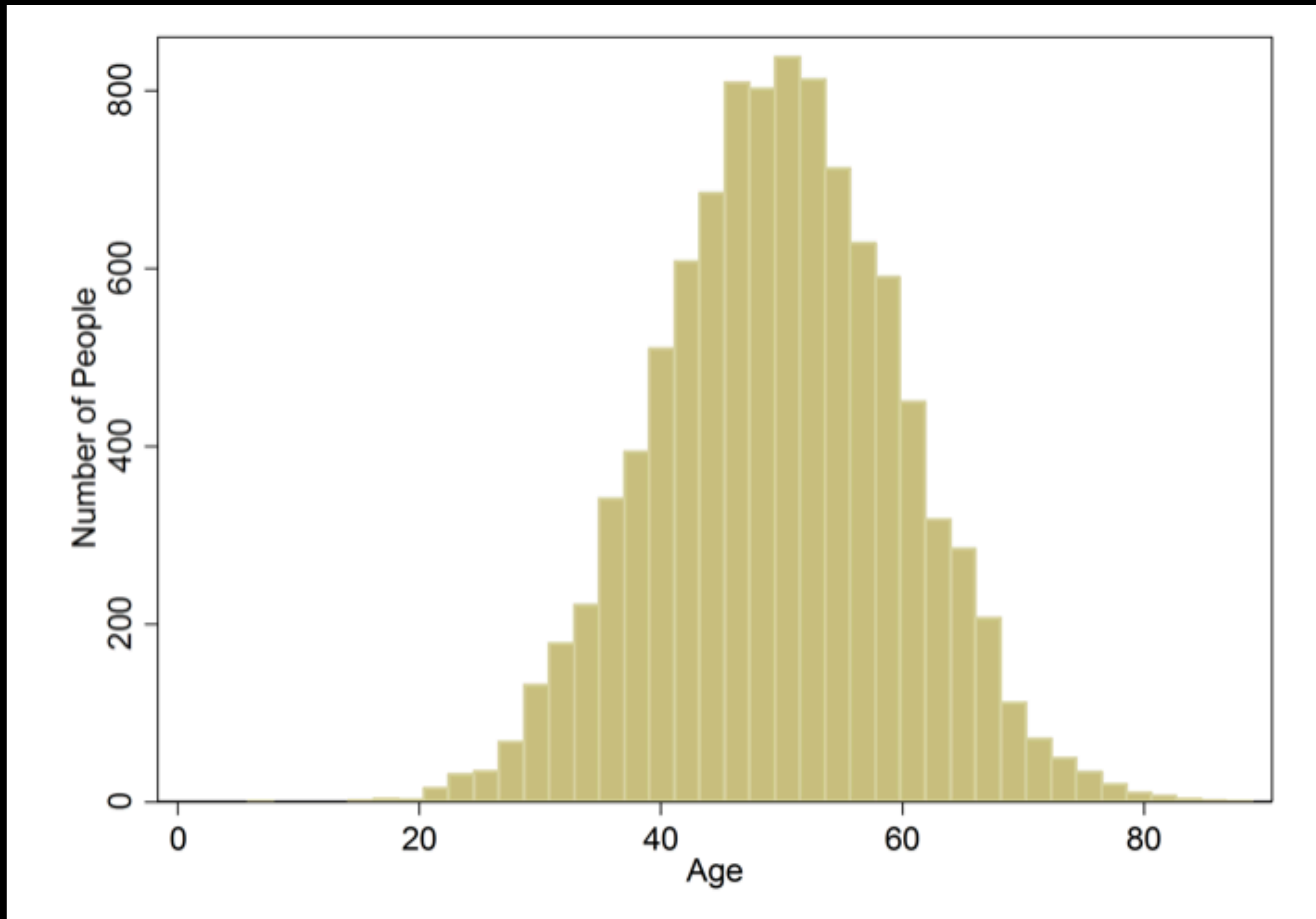
Later in you careers - in papers, for example - you will see this kind of things:

$$E[X]$$

"Expected value", the mean of a random variable

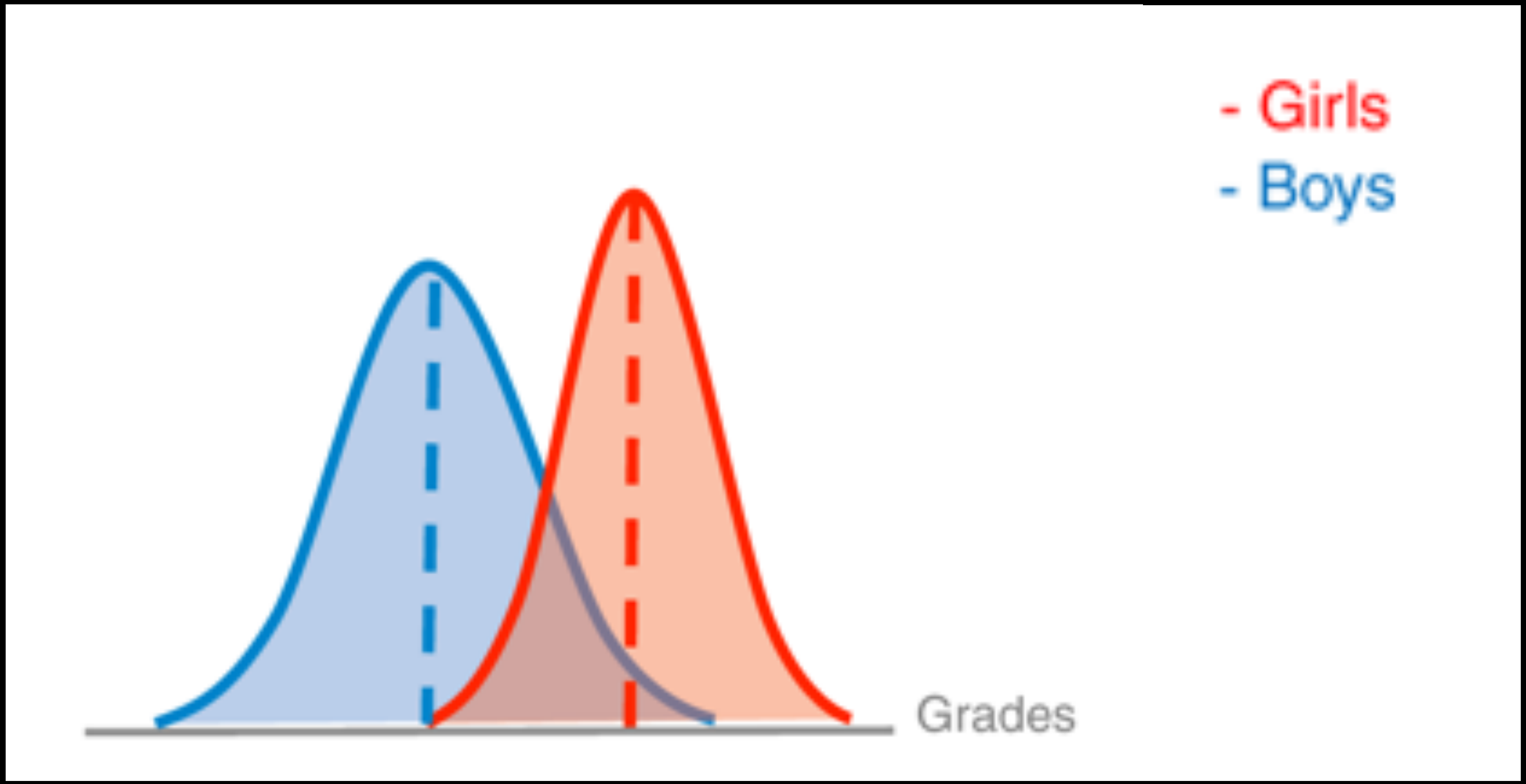
$$E[X] = \int xf(x)dx$$

Let’s see it in practice.
This is a “distribution”….

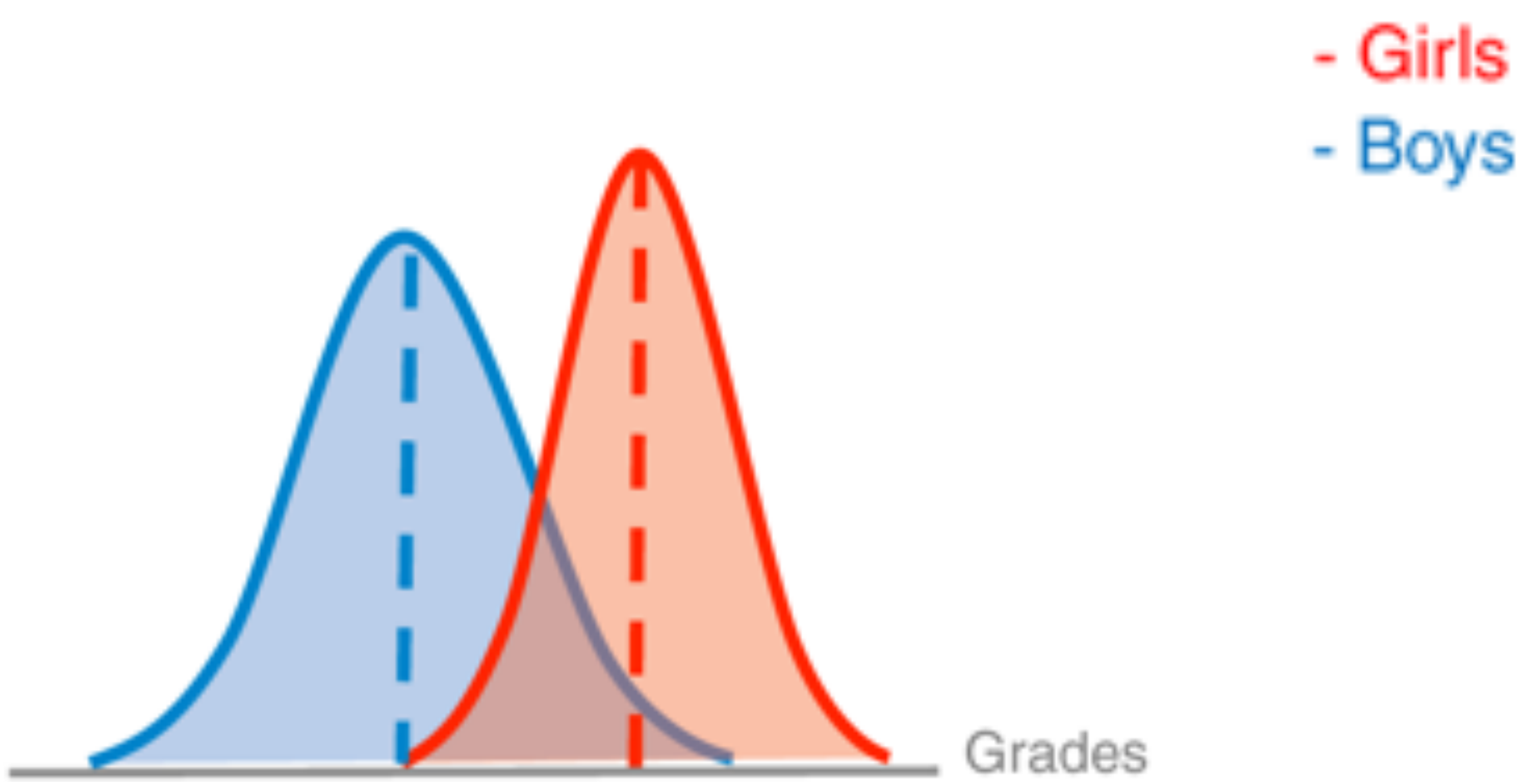


What would you say is the mean (or average) age here?

In most cases, the mean conveys a lot of information regarding a distribution

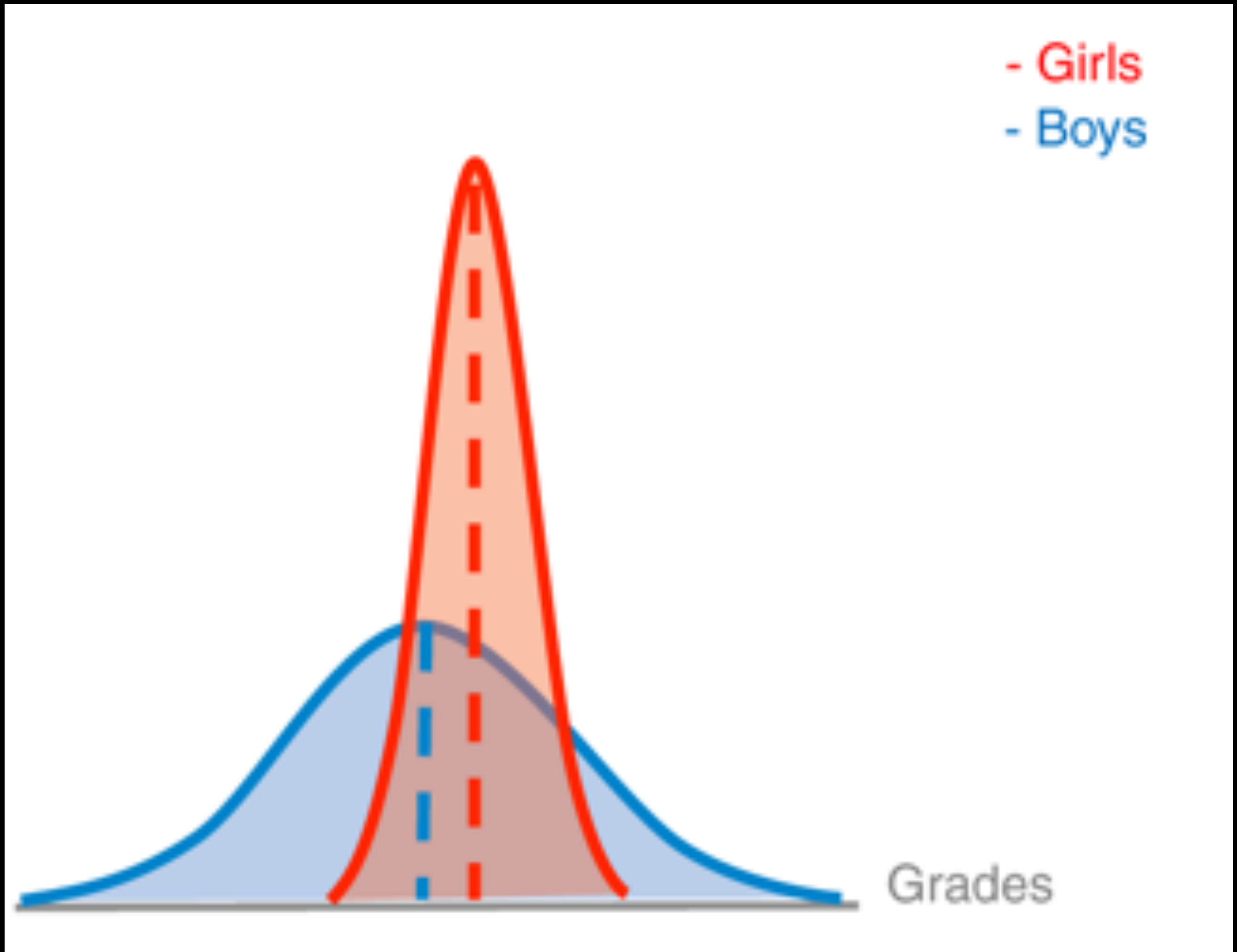


In most cases, the mean conveys a lot of information regarding a distribution

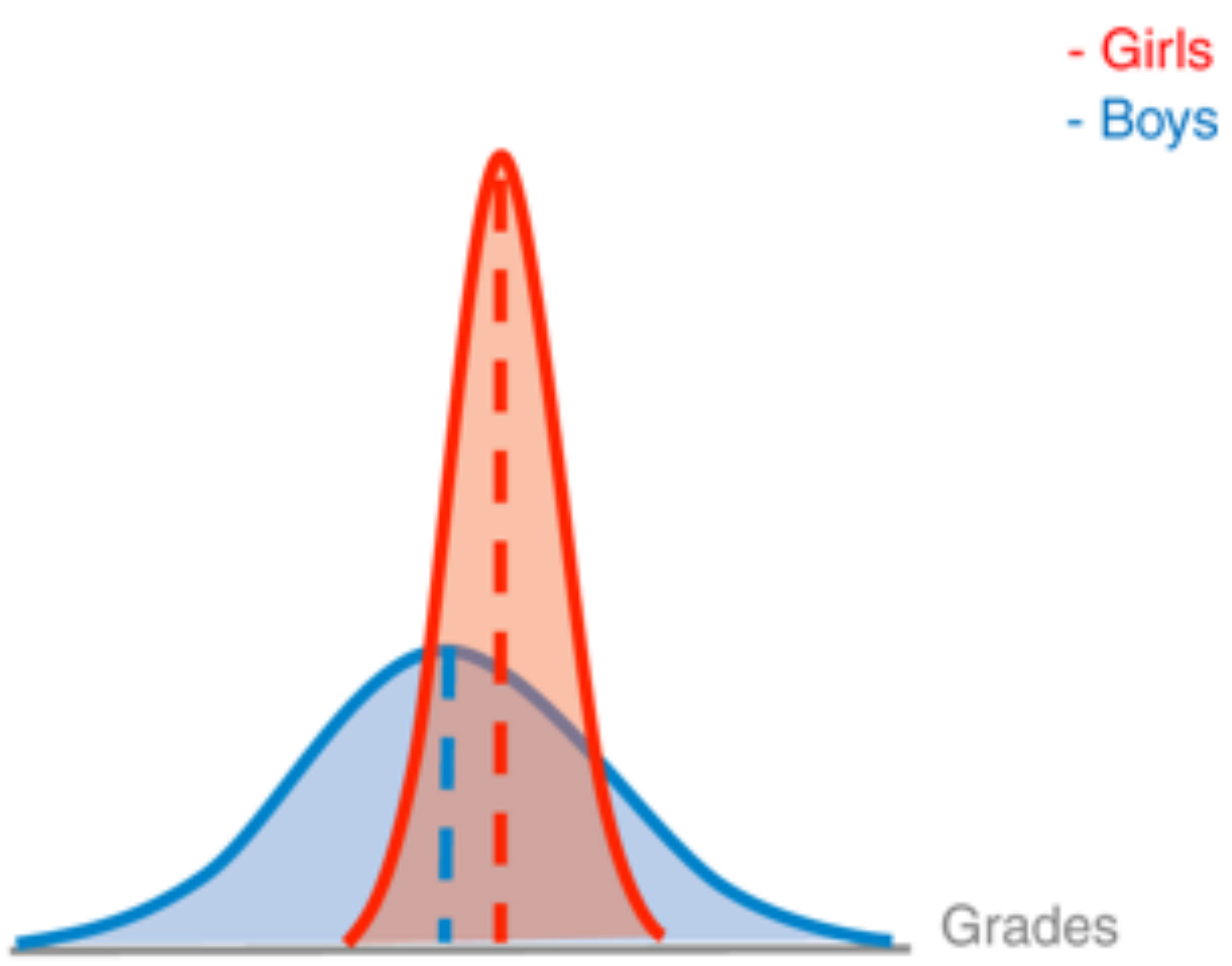


- On average girls grades much better,
- Girls grades similarly variable,
- More top scoring girls

In many cases it doesn't show the whole story

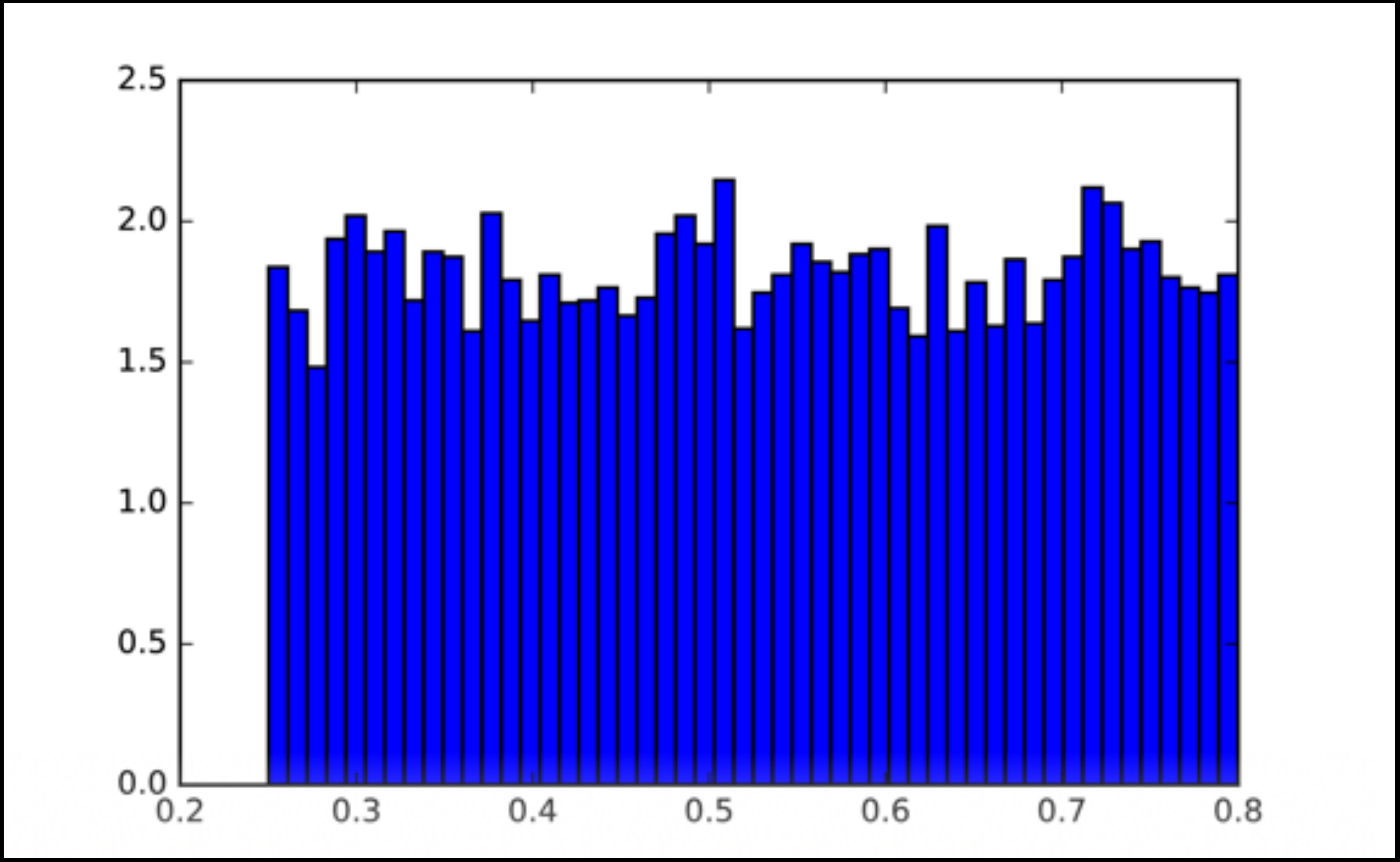


In many cases it doesn't show the whole story

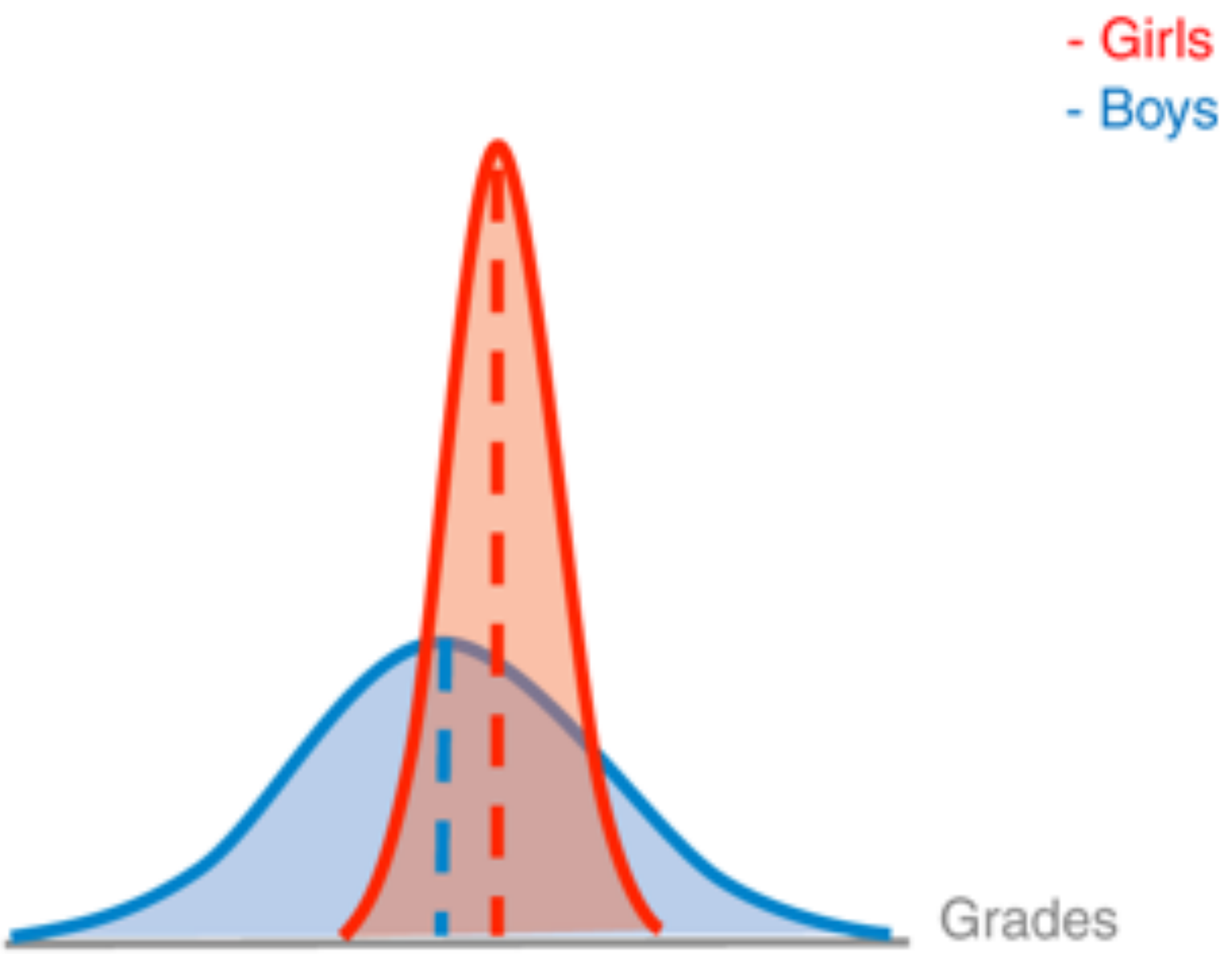


- On average girls grades slightly better,
- Girls grades much more consistent,
- Many fewer top scoring girls

In some cases it is not informative at all: uniform distribution



Dispersion
Variance and Standard Deviation



- On average girls grades slightly better,
- Girls grades much more consistent,
- Many fewer top scoring girls

Mean

$$X = \begin{bmatrix} 7 \\ 10 \\ 12 \\ 12 \\ 10 \\ 5 \\ 10 \end{bmatrix}$$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = 9.46$$

$$X - \bar{x} = \begin{bmatrix} (7 - 9.46) \\ (10 - 9.46) \\ (12 - 9.46) \\ (12 - 9.46) \\ (10 - 9.46) \\ (5 - 9.46) \\ (10 - 9.46) \end{bmatrix}$$

$$X - \bar{x} = \begin{bmatrix} -2.43 \\ 0.57 \\ 2.57 \\ 2.57 \\ 0.57 \\ -4.43 \\ 0.57 \end{bmatrix}$$

$$\sum_{i=1}^{N} (x_i - \bar{x}) = 0$$

$$X - \bar{x} = \begin{bmatrix} (7-9.46) \\ (10-9.46) \\ (12-9.46) \\ (12-9.46) \\ (10-9.46) \\ (5-9.46) \\ (10-9.46) \end{bmatrix} \qquad X - \bar{x} = \begin{bmatrix} -2.43 \\ 0.57 \\ 2.57 \\ 2.57 \\ 0.57 \\ -4.43 \\ 0.57 \end{bmatrix} \qquad \sum_{i=1}^{N} (x_i - \bar{x}) = 0$$

$$\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum \begin{bmatrix} (7-9.46)^2 \\ (10-9.46)^2 \\ (12-9.46)^2 \\ (12-9.46)^2 \\ (10-9.46)^2 \\ (5-9.46)^2 \\ (10-9.46)^2 \end{bmatrix} = \sum \begin{bmatrix} 5.9 \\ 0.33 \\ 6.6 \\ 6.6 \\ 0.33 \\ 19.6 \\ 0.33 \end{bmatrix} = 39.7$$

$$\sum_{i=1}^{N}(x_i-\bar{x})^2 = \sum \begin{bmatrix}(7-9.46)^2\\(10-9.46)^2\\(12-9.46)^2\\(12-9.46)^2\\(10-9.46)^2\\(5-9.46)^2\\(10-9.46)^2\end{bmatrix} = \begin{bmatrix}5.9\\0.33\\6.6\\6.6\\0.33\\19.6\\0.33\end{bmatrix} = 39.7$$

$$\frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N} = 39.7/7 = 5.67$$

This is the Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2}{N}$$

Let “y” be the deviation form the mean, squared

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

$$\sigma^2 = \bar{y}$$

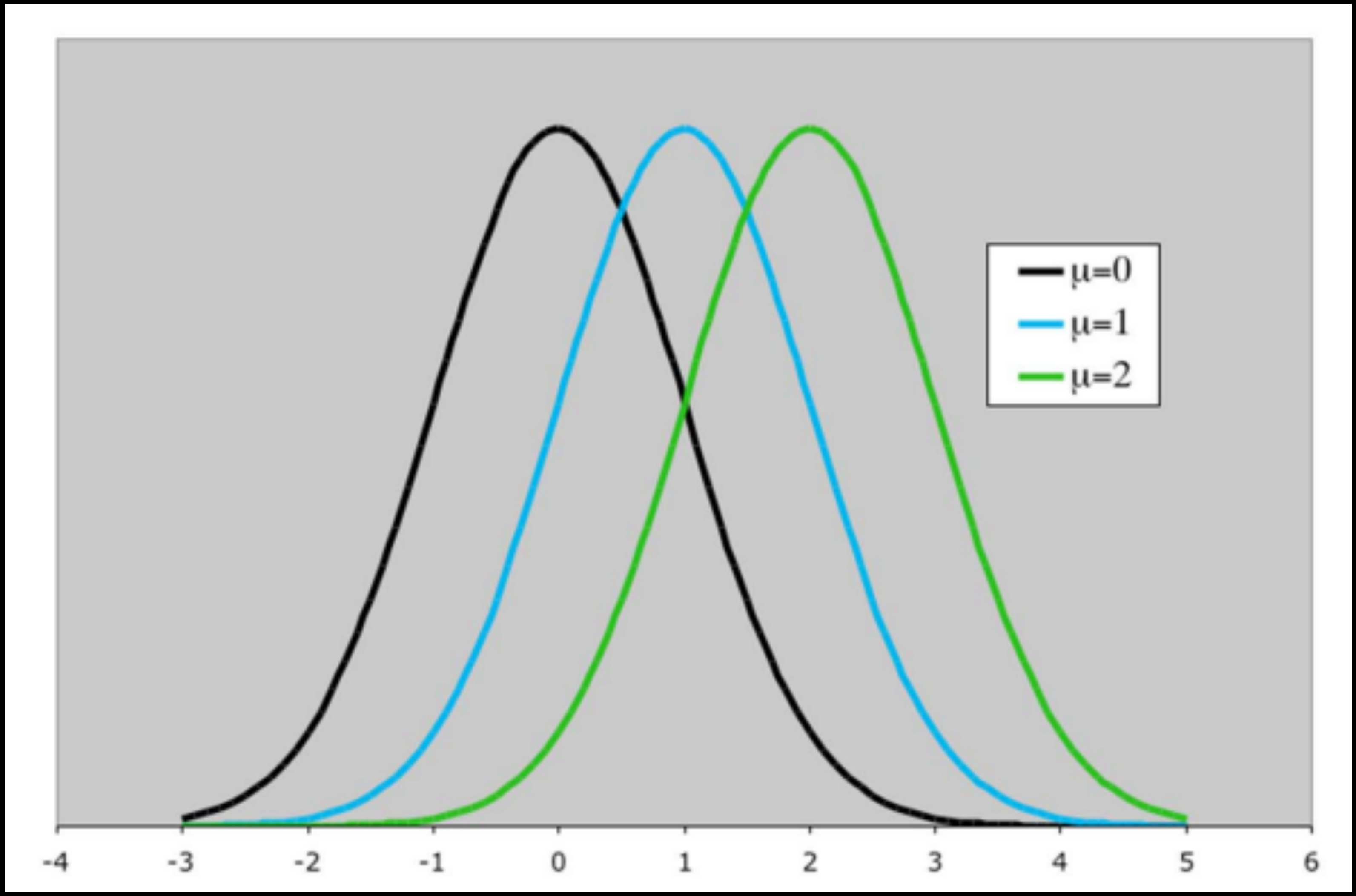
Variance is the average of the deviations form the mean, squared

By computing the “Standard deviation” we get thing back to scale

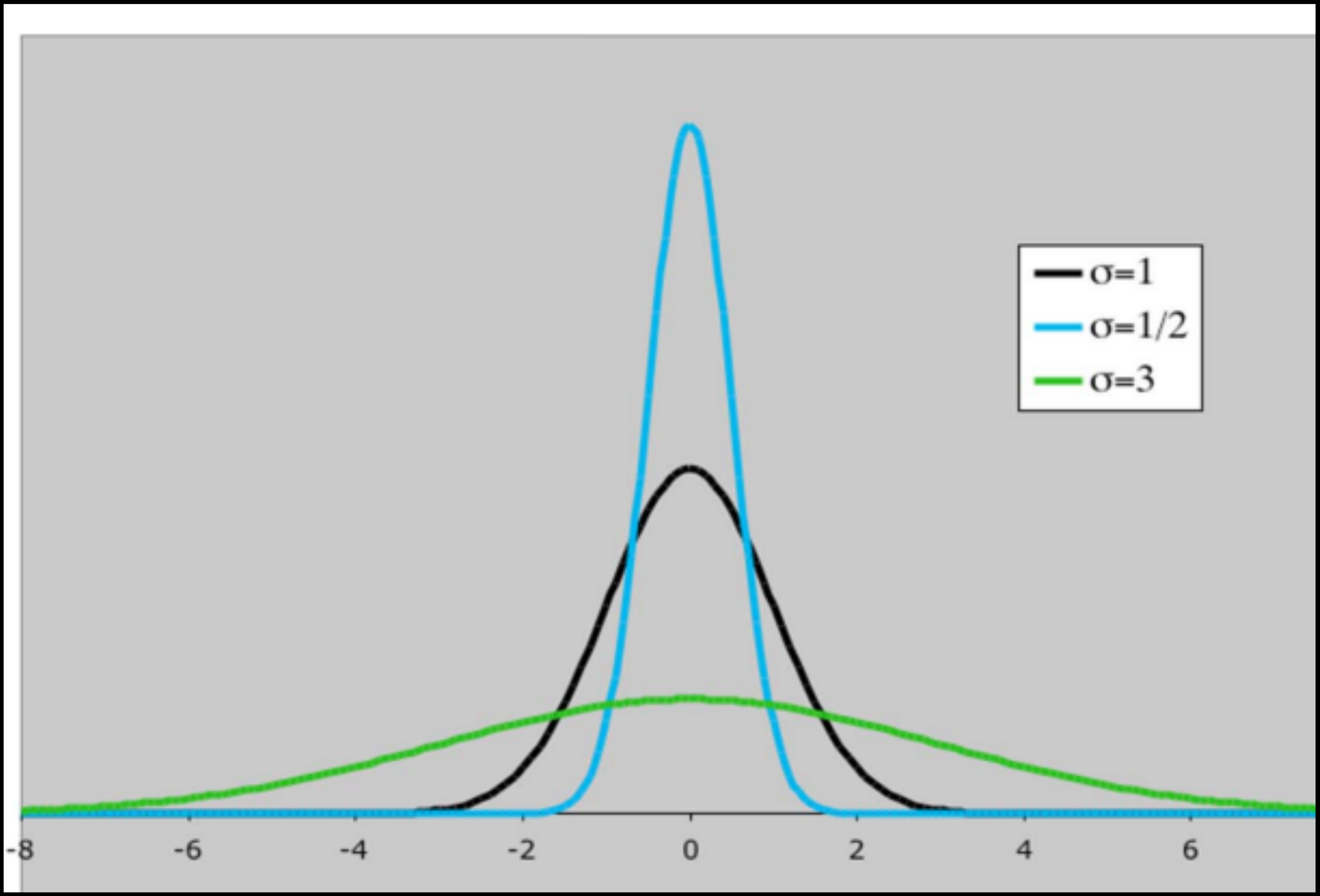
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N}} = 2.38$$

It computes the “average deviation” from the mean

Same variance but different means



Same mean but different variance



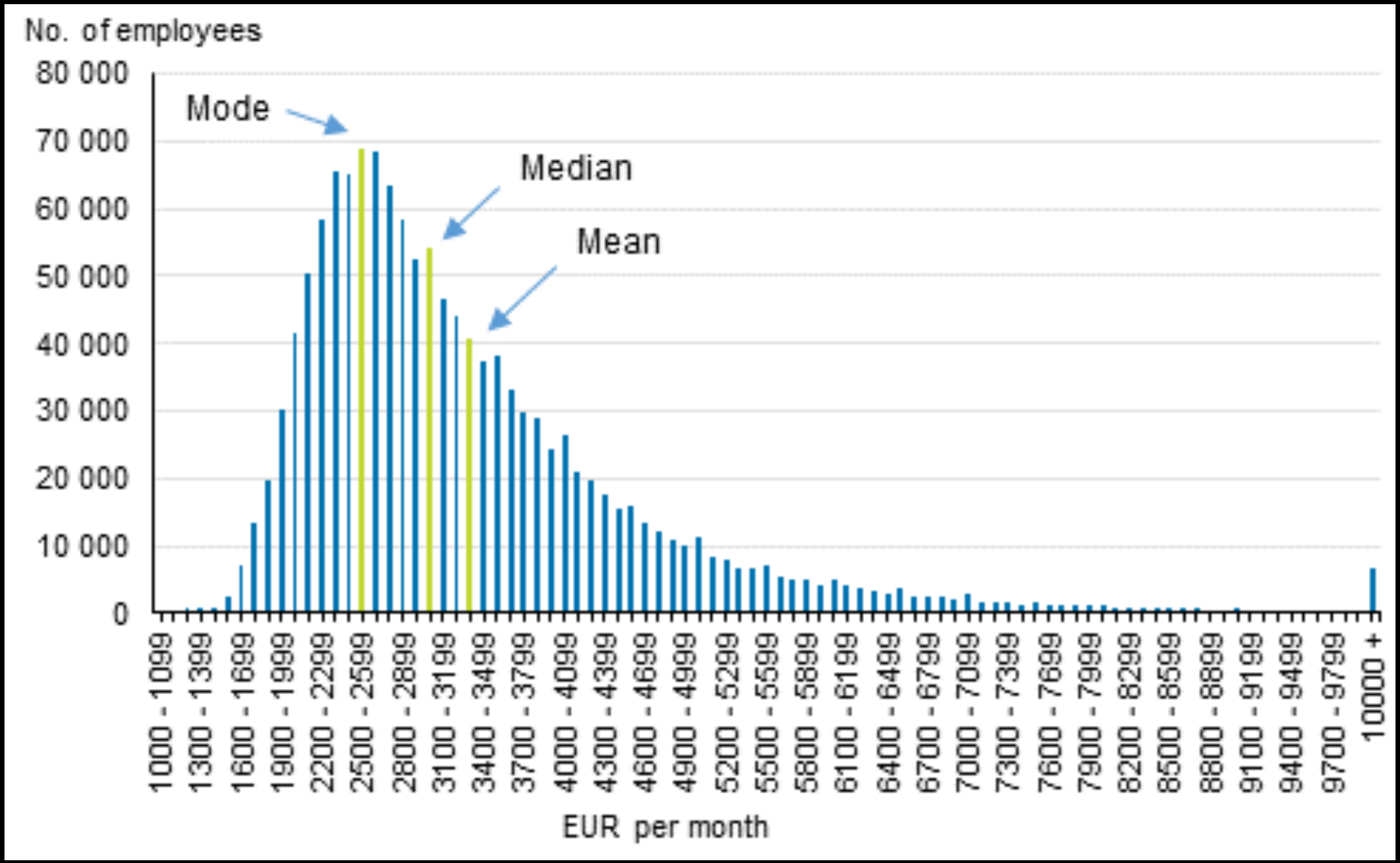
Something to think about …

Let "X", with $\bar{x} = \mu$

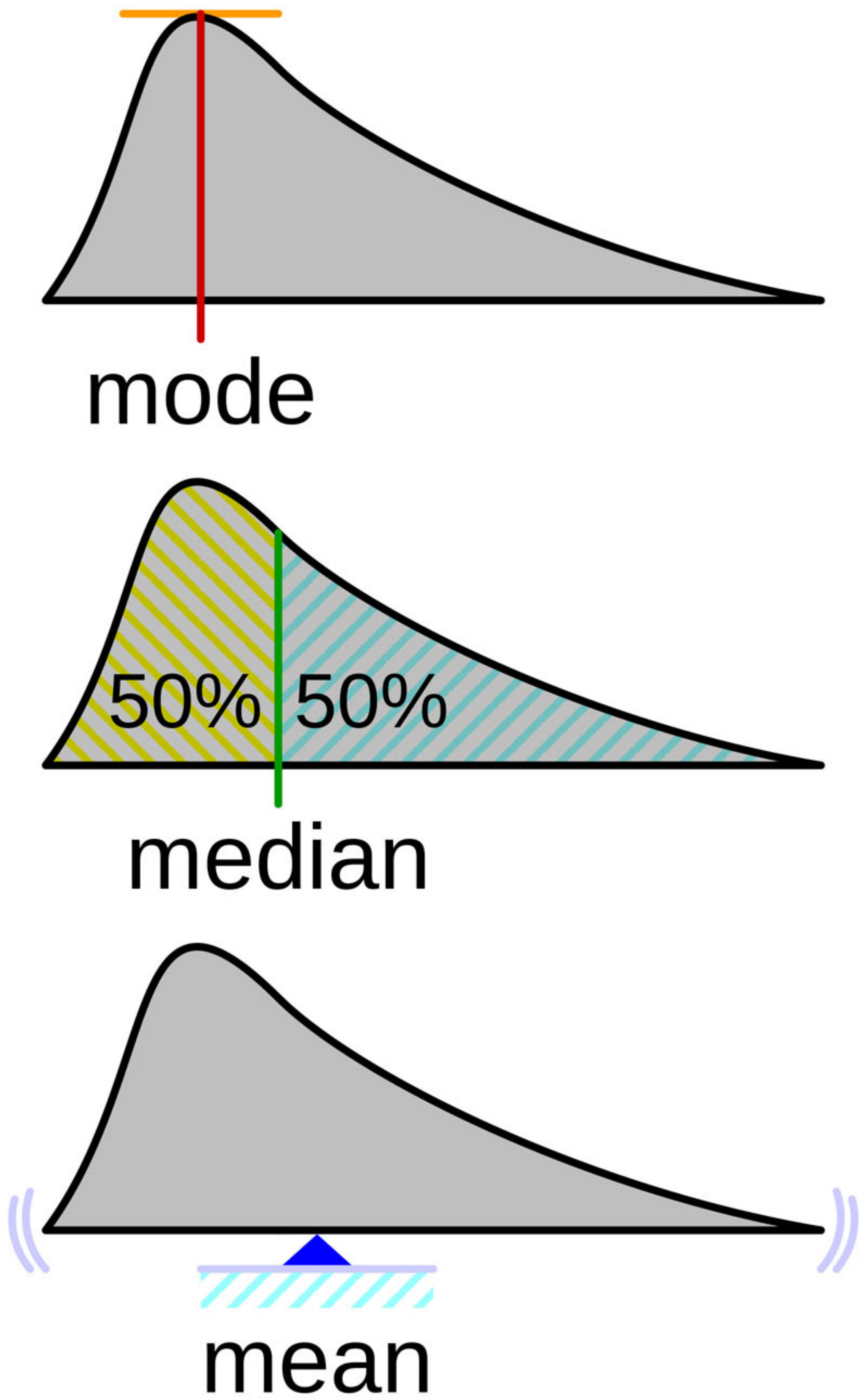
What is the mean of X+1 ?

What is the standard deviation of X+1 ?

Mode, Median, Quantiles



No. of employees
80 000
70 000
60 000
50 000
40 000
30 000
20 000
10 000
0
Mode
Median
Mean
1000 - 1099
1300 - 1399
1600 - 1699
1900 - 1999
2200 - 2299
2500 - 2599
2800 - 2899
3100 - 3199
3400 - 3499
3700 - 3799
4000 - 4099
4300 - 4399
4600 - 4699
4900 - 4999
5200 - 5299
5500 - 5599
5800 - 5899
6100 - 6199
6400 - 6499
6700 - 6799
7000 - 7099
7300 - 7399
7600 - 7699
7900 - 7999
8200 - 8299
8500 - 8599
8800 - 8899
9100 - 9199
9400 - 9499
9700 - 9799
10000 +
EUR per month


mode
50% 50%
median
mean

Mode

$$X = \begin{bmatrix} 7 \\ 10 \\ 12 \\ 12 \\ 10 \\ 5 \\ 10 \end{bmatrix}$$

Mode(X) = ?

“The most frequent value of X”

Mode(X) = 10

Median

"a" is the median of X if 50% or more of observations take on values equal or lower than "a"

$$P(X \leq a) \geq 0.5$$

$$X = \begin{bmatrix} 7 \\ 10 \\ 12 \\ 12 \\ 10 \\ 5 \\ 10 \end{bmatrix}$$

X	Freq.	Percent	Cum.
5	1	14.29	14.29
7	1	14.29	28.57
10	3	42.86	71.43
12	2	28.57	100.00
Total	7	100.00	

Median(X) = 10

X	Freq.	Percent	Cum.
.043255	1	3.23	3.23
.066943	1	3.23	6.45
.0722964	1	3.23	9.68
.0880495	1	3.23	12.90
.1288747	1	3.23	16.13
.1327082	1	3.23	19.35
.1359006	1	3.23	22.58
.1749891	1	3.23	25.81
.1778325	1	3.23	29.03
.2468877	1	3.23	32.26
.2775184	1	3.23	35.48
.2977743	1	3.23	38.71
.3325624	1	3.23	41.94
.3764437	1	3.23	45.16
.4087105	1	3.23	48.39
.4242016	1	3.23	51.61
.4476188	1	3.23	54.84
.4675523	1	3.23	58.06
.5160881	1	3.23	61.29
.5346152	1	3.23	64.52
.5536171	1	3.23	67.74
.5662265	1	3.23	70.97
.6794177	1	3.23	74.19
.6817465	1	3.23	77.42
.7124024	1	3.23	80.65
.7551366	1	3.23	83.87
.7677861	1	3.23	87.10
.7767794	1	3.23	90.32
.8302209	1	3.23	93.55
.8745816	1	3.23	96.77
.9339853	1	3.23	100.00
Total	31	100.00	

Median(X) = 0.42

Percentiles

"a" is the "p" percentile of X if p% or more of observations take on values equal or lower than "a"

$$P(X \leq a) \geq p/100$$

The median is the percentile 50:

$$P(X \leq a) \geq 50/100 = 0.5$$

$$X = \begin{bmatrix} 7 \\ 10 \\ 12 \\ 12 \\ 10 \\ 5 \\ 10 \end{bmatrix}$$

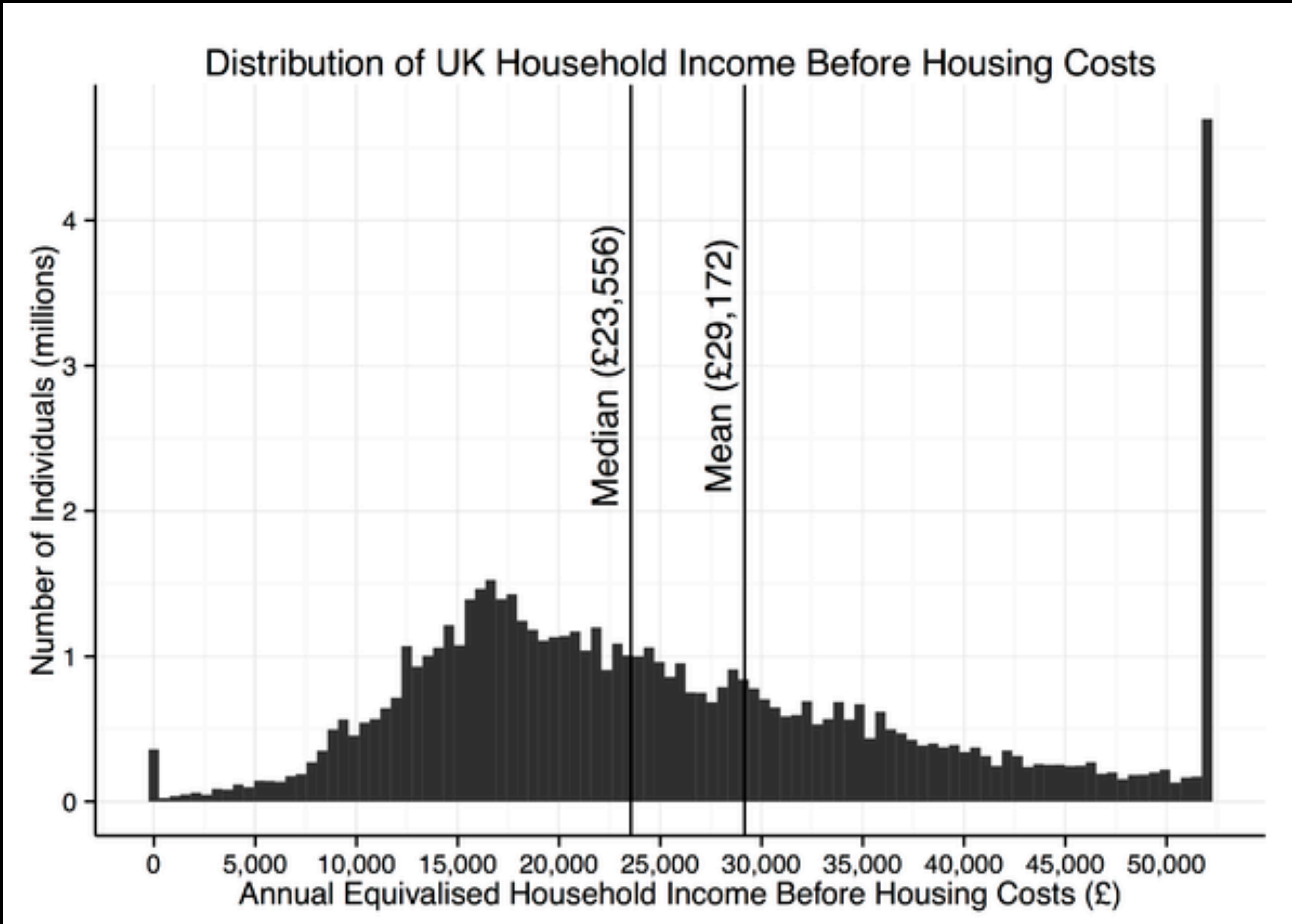
X	Freq.	Percent	Cum.
5	1	14.29	14.29
7	1	14.29	28.57
10	3	42.86	71.43
12	2	28.57	100.00
Total	7	100.00	

median(X) = p50(X) = 10

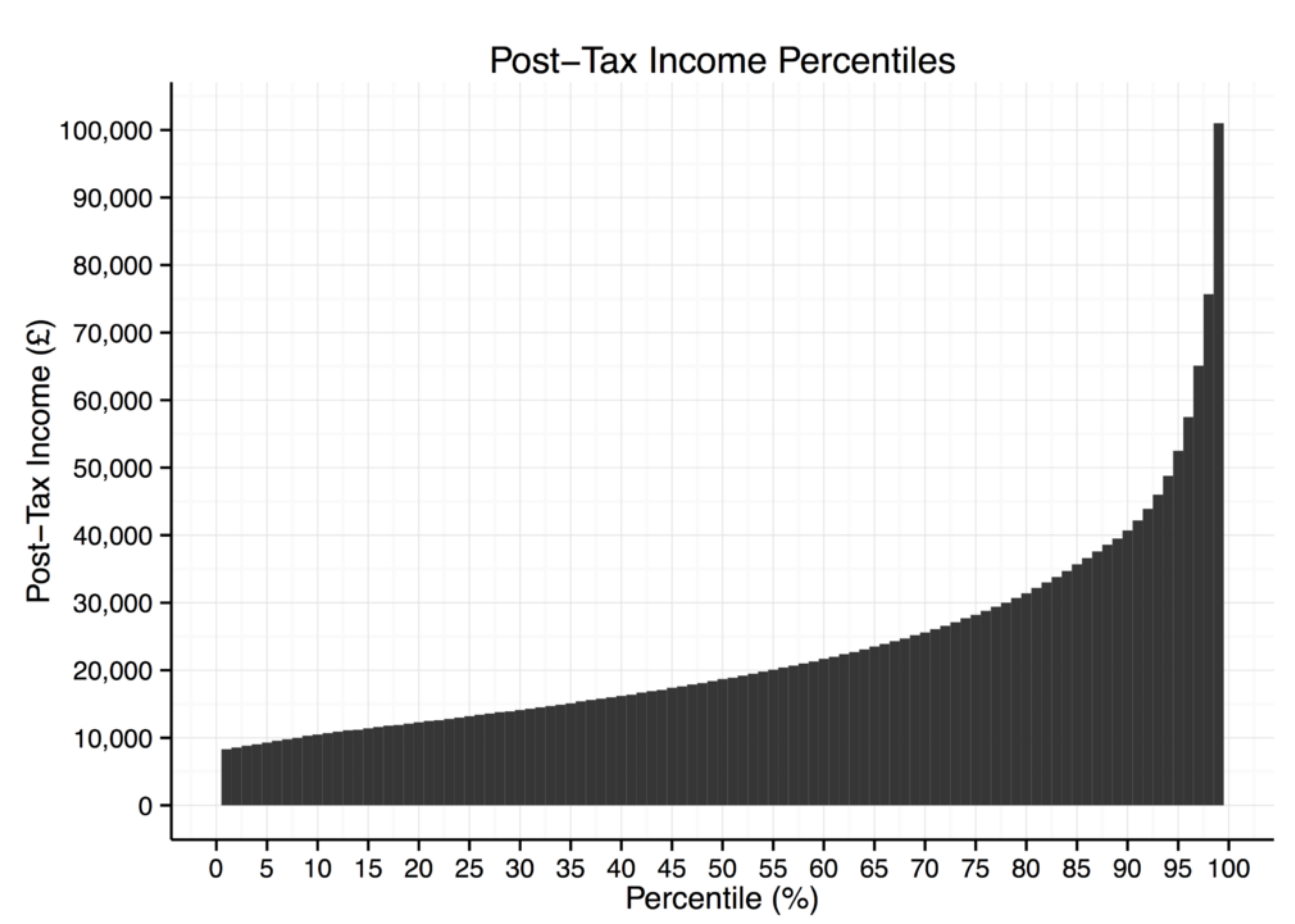
X	Freq.	Percent	Cum.
.043255	1	3.23	3.23
.066943	1	3.23	6.45
.0722964	1	3.23	9.68
.0880495	1	3.23	12.90
.1288747	1	3.23	16.13
.1327082	1	3.23	19.35
.1359006	1	3.23	22.58
.1749891	1	3.23	25.81
.1778325	1	3.23	29.03
.2468877	1	3.23	32.26
.2775184	1	3.23	35.48
.2977743	1	3.23	38.71
.3325624	1	3.23	41.94
.3764437	1	3.23	45.16
.4087105	1	3.23	48.39
.4242016	1	3.23	51.61
.4476188	1	3.23	54.84
.4675523	1	3.23	58.06
.5160881	1	3.23	61.29
.5346152	1	3.23	64.52
.5536171	1	3.23	67.74
.5662265	1	3.23	70.97
.6794177	1	3.23	74.19
.6817465	1	3.23	77.42
.7124024	1	3.23	80.65
.7551366	1	3.23	83.87
.7677861	1	3.23	87.10
.7767794	1	3.23	90.32
.8302209	1	3.23	93.55
.8745816	1	3.23	96.77
.9339853	1	3.23	100.00
Total	31	100.00	

p30(X) = 0.2468

p90(X) = 0.8302



Distribution of UK Household Income Before Housing Costs
Median (£23,556)
Mean (£29,172)
Number of Individuals (millions)
0
1
2
3
4
0
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
Annual Equivalised Household Income Before Housing Costs (£)


Post−Tax Income Percentiles
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Post−Tax Income (£)
0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percentile (%)

Descriptive Statistics - discrete variables

Dichotomous case: proportion is the only parameter

$$X = \begin{bmatrix} Male \\ Male \\ Female \\ Male \\ Male \\ Male \\ Female \end{bmatrix}$$

$$p = P(Male)$$

$$q = P(Female) = 1 - p$$

gender	Freq.	Percent	Cum.
Female	2	28.57	28.57
Male	5	71.43	100.00
Total	7	100.00	

Proportion can be also expressed as a mean

$$X = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{(1+1+0+1+1+1+0)}{7} = 0.71 = p$$

$$q = 1 - p = 1 - \bar{x}$$

Proportion can be also expressed as a mean

$$X = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{(1+1+0+1+1+1+0)}{7} = 0.71 = p = P(1) = P(Male)$$

$$q = 1 - p = 1 - \bar{x}$$

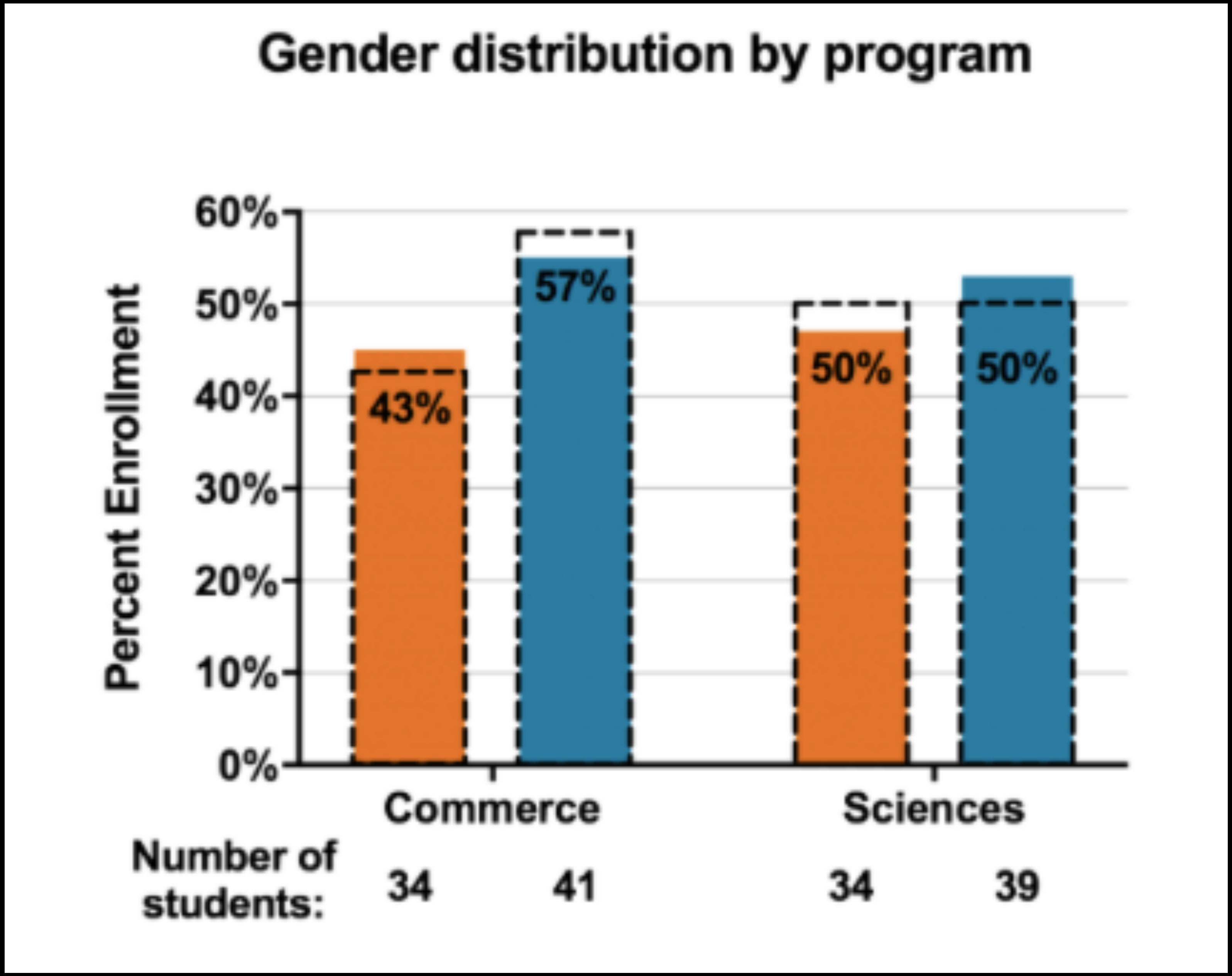
Why does this work? Recall that

$$\bar{x} = \sum_{\text{All x}} x = P(x) * x$$

Weighted average

$$\bar{x} = P(1) * 1 + P(0) * 0 = p * 1 + (1-p) * 0 = p$$

Dispersion? Variance/SD ?



Variance

$$X = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

$$\sigma^2 = p(1-p)$$

$$\sigma = \sqrt{p(1-p)}$$

$$(X-p)^2 = \begin{bmatrix} 0.08 \\ 0.08 \\ 0.51 \\ 0.08 \\ 0.08 \\ 0.82 \\ 0.51 \end{bmatrix}$$

$$P[(1-p)^2] = p$$

$$P[(0-p)^2] = 1-p$$

Why?

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N} = \overline{(x_i - \bar{x})^2} =$$

$$\sigma^2 = P[(1-p)^2] * (1-p)^2 + P[(0-p)^2] * (0-p)^2$$

$$= p * (1-p)^2 + (1-p) * (0-p)^2$$

$$= p * (1 - 2p + p^2) + (1-p) * p^2$$

$$= p - 2p^2 + p^3 + p^2 - p^3$$

$$= p - p^2 = p(1-p)$$

Many valued discrete variables:

$$X = \begin{bmatrix} A \\ B \\ C \\ A \\ B \\ C \\ A \end{bmatrix}$$

The number of parameters for a variable with k categories is k-1

Here k=3, so we need k-1=2 parameters

$$P[A] = p = 3/7$$

$$P[B] = q = 2/7$$

$$P[C] = 1 - (p + q) = 1 - 5/7 = 2/7$$

Variance

$$\sigma^2 = p * q * (1 - p - q)$$

SD

$$\sigma = \sqrt{p * q * (1 - p - q)}$$