2 - Summary parameters [ES 2.4], [PS 1.3]

A **central tendency** approximates a dataset or a data vector $\vec{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ by a single number c.

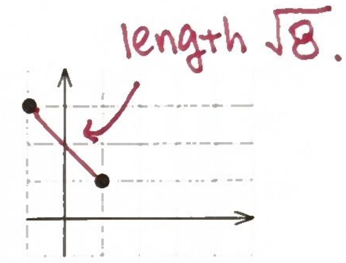
Here $\mathbb{R}^n$ is the set of n-tuples of real numbers aka the n-dimensional space of real numbers.

More precisely, a **central tendency** is the homogeneous vector $(c, c, \ldots, c)$ closest to $\vec{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ using some notion of distance.

Definition. The **Euclidean distance** between vectors $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $\vec{y} = (y_1, y_2, \ldots, y_n)$ is

$$||\vec{x} - \vec{y}|| = \left(\sum_{i=1}^{n} (x_i - y_i)^2 \right)^{1/2}.$$

Example 1. Find the Euclidean distance between $(1, 1)$ and $(-1, 3)$.



$$||(1,1)-(-1,3)|| = ||(2,-2)|| = \sqrt{2^2+2^2} = \sqrt{8}$$

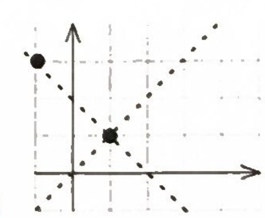
Theorem. The mean of $\vec{x} = (x_1, x_2, \ldots, x_n)$ is the number c which minimizes the distance between $\vec{x} = (x_1, x_2, \ldots, x_n)$ and $(c, c, \ldots, c)$.

Why? To minimize $f(c) = ||\vec{x} - \vec{c}||^2 = \sum_{i=1}^{n} (x_i - c)^2$, we set $f'(c) = 0$:

$$\Rightarrow \quad 0 = f'(c) \underset{\text{chain rule}}{=\!=\!=} \sum_{i=1}^{n} -2 \cdot (x_i - c)$$

$$\Rightarrow \quad 0 = \sum_{i=1}^{n} (x_i - c) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} c = \sum_{i=1}^{n} x_i - n \cdot c \Rightarrow \boxed{c = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

Example 1 shows how $\vec{x}$ decomposes into the sum of two perpendicular lengths: the mean part $\vec{\mu}$ and the remaining part where the data values "varies" about the mean:



$$\vec{x} = \vec{\mu} + (\vec{x} - \vec{\mu}) \quad \Longrightarrow \quad ||\vec{x}||^2 = ||\vec{\mu}||^2 + ||(\vec{x} - \vec{\mu})||^2 \quad \text{(Pythagorean theorem)}$$

Definition. The **population variance** of $\vec{x}$ is

$$\text{pop.var}(\vec{x}) = \sigma^2 = \frac{1}{n} ||\vec{x} - \vec{\mu}||^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_i)^2$$

The **population standard deviation** of $\vec{x}$ is its square root

$$\text{pop.sd}(\vec{x}) = \sigma = \frac{1}{\sqrt{n}} ||\vec{x} - \vec{\mu}|| = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_i)^2}$$

Example 2. Find the population variance of $\vec{x} = (-1, 3)$.

$$\mu = \frac{-1+3}{2} = 1 \Rightarrow \sigma^2 = \frac{1}{2} ||(-1,3) - (1,1)||^2 = \frac{1}{2} ||(-2,2)||^2 = \frac{1}{2} [2^2 + 2^2] = \boxed{4}$$

Example 3. Find the population variance of $\vec{x} = (-1, 3, -1, 3)$.

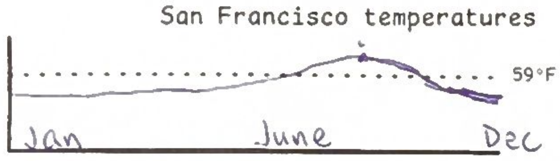
$\mu = 1$ again so $\sigma^2 = \frac{1}{4}\|(-1,3,-1,3)-(1,1,1,1)\|^2 = \frac{1}{4}\|(-2,2,-2,2)\|^2 = \frac{1}{4}\cdot 4 \cdot 2^2 = \boxed{4}$

Moral: Dividing by n ensures the standard deviation does not increase for silly reasons just by collecting more data.

Example 4. Shown are average monthly temperatures. Where would you rather live?

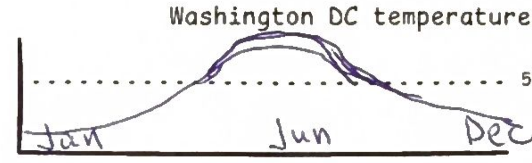
°F	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	μ
San Francisco	52	54	55	56	58	60	60	65	66	62	57	53	59
Washington DC	38	40	48	58	67	76	81	79	72	61	50	42	59

San Francisco temperatures



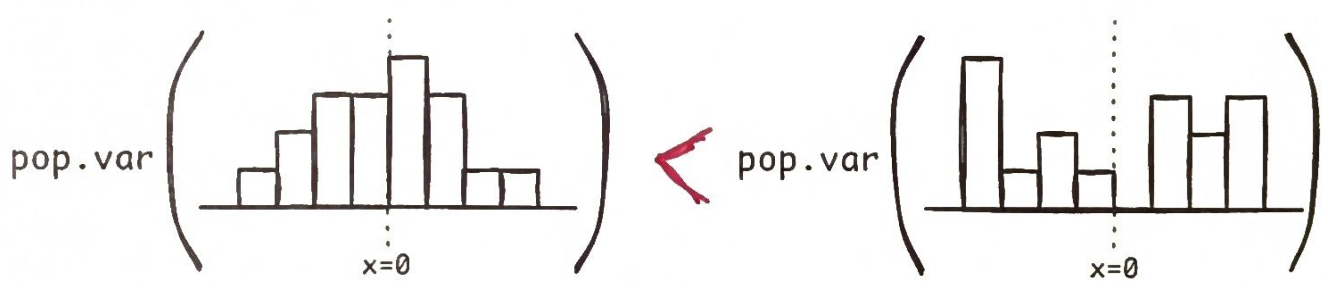
less variance, more consistent

Washington DC temperatures



more variance, less consistent temperature

Example 5. (Variance in histogram)



Both histograms have mean 0, but the right histogram has larger variance.

Exercise. Find the population variance of the data set (0, 5, 10).

$\mu = \frac{1}{3}(0+5+10) = 5 \Rightarrow \sigma^2 = \frac{1}{3}\|(-5,0,5)\|^2 = \frac{2\cdot 5^2}{3} = \boxed{\frac{50}{3}}$

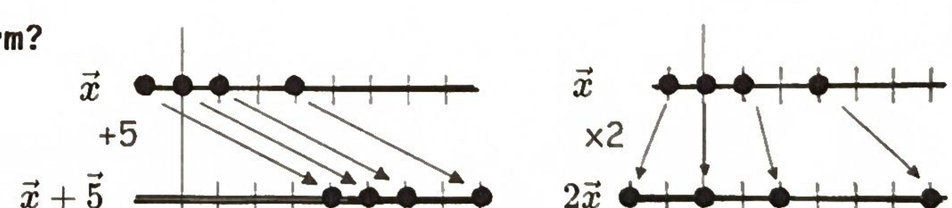
How do mean and variance transform?

$\text{mean}(\vec{x}+c) = \text{mean}(\vec{x}) + c$

$\text{pop.stdev}(\vec{x}+c) = \text{pop.stdev}(\vec{x})$

$\text{mean}(c\vec{x}) = c \cdot \text{mean}(\vec{x})$

$\text{pop.stdev}(c\vec{x}) = c \cdot \text{pop.stdev}(\vec{x})$



Example 6. Monthly temperatures in Glassboro have mean 55°F and population standard deviation 18°F. Re-express these numbers in Celsius.

μ: $°C = \frac{5}{9}(°F - 32) = \frac{5}{9}(55°F - 32) = \frac{5}{9}(23) \approx \boxed{13\,°C}$

σ: $°F \xrightarrow{-32} °F - 32 \xrightarrow{\times \frac{5}{9}} \frac{5}{9}(°F - 32) = °C$

$18 \xrightarrow[\text{on standard dev.}]{\text{no effect}} 18 \xrightarrow{\times\frac{5}{9}} 18 \cdot \frac{5}{9} = \boxed{10\,°C}$