3 - Position in distribution [ES 2.5, PS 2.1, 2.2]

Normal distribution and Empirical Rule [ES 2.4, PS 2.1]

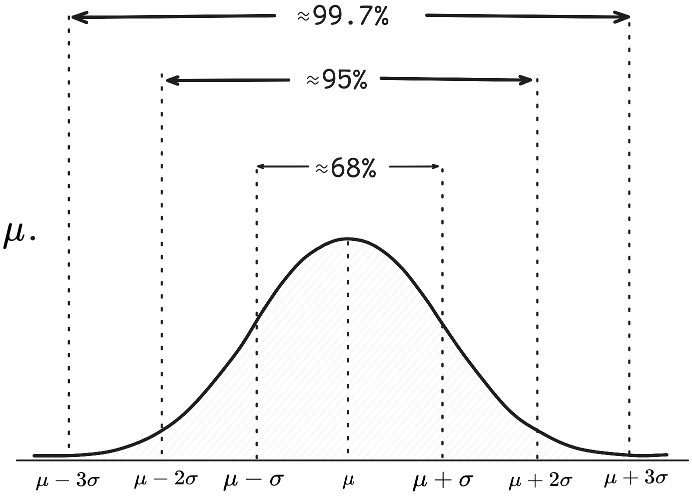
Many datasets naturally a bell-shaped distribution called the **normal distribution** whose shape is completely determined by the mean and the standard deviation.
Examples:
- scores or tests taken by many people (SAT exams, IQ tests)
- repeated careful measurements of the same quantity
- characteristics of many biological populations (cricket length, corn yields)

68-95-99.7 Rule.
In a normal distribution with mean μ and standard deviation σ:

≈68% of the data fall within σ of the mean μ.

≈95% of the data fall within 2σ of μ.

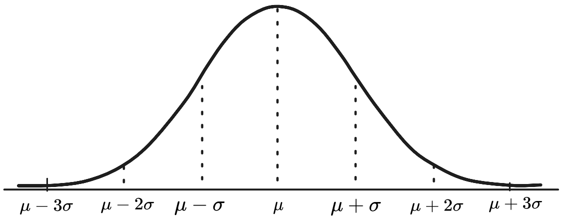
≈99.7% of the data fall within 3σ of μ.



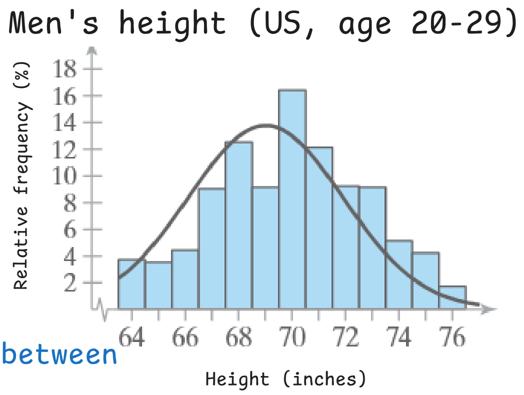
Example 1. NCHS survey finds women (US, age 20-29) have mean height of 64.2 inches and standard deviation of 2.9 inches. Estimate the percent of women whose heights are between 58.4–64.2 inches.

Answer. Distribution of women's heights is roughly bell-shaped.
Also, 58.4 inches = mean - 2 pop.sd
So we want the region from $\mu-2\sigma$ to μ.

This is 95/2 = 47.5% of women by the 68-95-99.7 rule.



Example 2. Men (US, age 20-29) heights follow a rough bell-shape with mean 69.4 inches and standard deviation of 2.9 inches.
(a) Estimate the two heights containing the middle 95% of the data.
(b) Is a 25 year old with height 74 inches unusual?



Answer(a): By 68-95-99.7 rule, 95% of data are between

$$\mu+2\sigma=69.4+2\times2.9=75.2\text{ inches}$$

$$\mu-2\sigma=69.4-2\times2.9=63.6\text{ inches}$$

Standard normal distribution, z-score [ES 2.5, PS 2.1]

We can standardize datasets to compare them.

$$\begin{tikzcd} \vec{x} \arrow[r, "\text{Centering}"] & \vec{x}-\mu \arrow[r, "\text{Standardizing}"] & \frac{1}{\sigma}(\vec{x}-\mu) \\ \text{Original dataset} & \text{Mean = 0} & \begin{array}{c}\text{Mean = 0}\\ \text{pop.sd = 1}\end{array} \end{tikzcd}$$

The **standard score** or **z-score** of a data point x is $z = \dfrac{x-\mu}{\sigma}$.

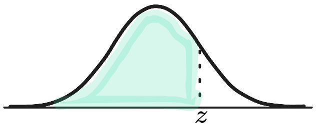
Example 3. The dataset $\vec{x} = (x_1, x_2)$ = (-1, 3) has mean 1 and pop.sd 2. So its standardization is

$$\frac{1}{\sigma}(\vec{x}-\mu) = \frac{1}{2}[(-1,3)-(1,1)] = \frac{1}{2}(-2,2) = (\underset{\boxed{\text{z-score of } x_1}}{-1}, \underset{\boxed{\text{z-score of } x_2}}{1})$$

Standardizing a normal distribution gives the **standard normal distribution**.

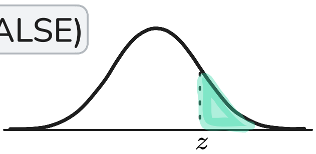
`pnorm(z)`

finds the area to the left of z.



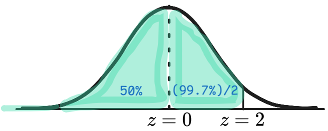
`pnorm(z, lower.tail=FALSE)`

finds the area to the right of z.



Example 4. Estimate pnorm(2) without electronics.

In the standard normal distribution,
z=2 is 2 standard deviations out.
The area to the left of z=0 is 50% of area.
The area from z=0 to z=2 is (95)/2 = 47.5 by the 68-95-99.7 rule.
Answer: 50% + 47.5% = 97.5% or 0.975.



Boxplot, percentiles, IQR, outliers [ES 2.5]

We visualize general datasets with boxplots.

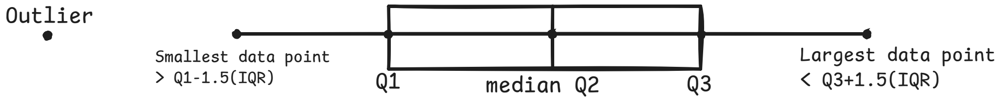
- pth percentile = p % of data is below this value
- First quartile (Q1) = 25 percentile
- Second quartile (Q2) = 50 percentile (median)
- Third quartile (Q3) = 75 percentile

Example 5. Find Q1, Q2, Q3 of amount (gallons/year) of fuel wasted in 15 largest US urban areas:

11 20 22 23 24 25 25 25 28 29 29 30 33 35 35

Definition. The interquartile range (IQR) of a dataset is Q3-Q1.
A datapoint is an outlier if it is > Q3 + 1.5(IQR) or < Q1 - 1.5(IQR)

Draw boxplot like so:



Example 6. For Example 5, draw its boxplot and describe distribution.

IQR = 30 - 23 = 7, Q3 + 1.5(IQR) = 30 + 10.5 = 40.5, Q1 - 1.5(IQR) = 12.5.
So 11 is an outlier.