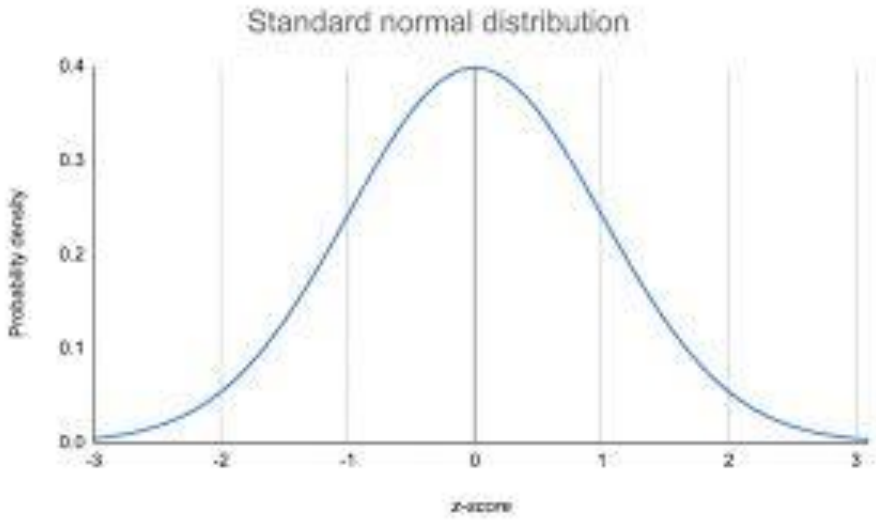
Normal Distribution

What is it?

- The normal distribution is a function that defines how a set of measurements is distributed around the center of these measurements (i.e., the mean).
- Many natural phenomena in real life can be approximated by a bell-shaped frequency distribution
- Also known as the Gaussian distribution.
- N(mean,sd)

Properties

- Bell-shaped, unimodal and symmetric distribution
- A unimodal distribution is a distribution with **one clear peak** or most frequent value.
- Curve extends to –inf to +inf
- Area under the curve is 1



Standard deviation

- The unit of measurement usually given when talking about statistical significance is the standard deviation, expressed with the lowercase Greek letter sigma (σ).
- The term refers to the amount of variability in a given set of data: whether the data points are all clustered together, or very spread out.
- A low standard deviation indicates that the values tend to be close to the mean of the set.
- A high standard deviation indicates that the values are spread out over a wider range.

Standard deviation

Formula

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

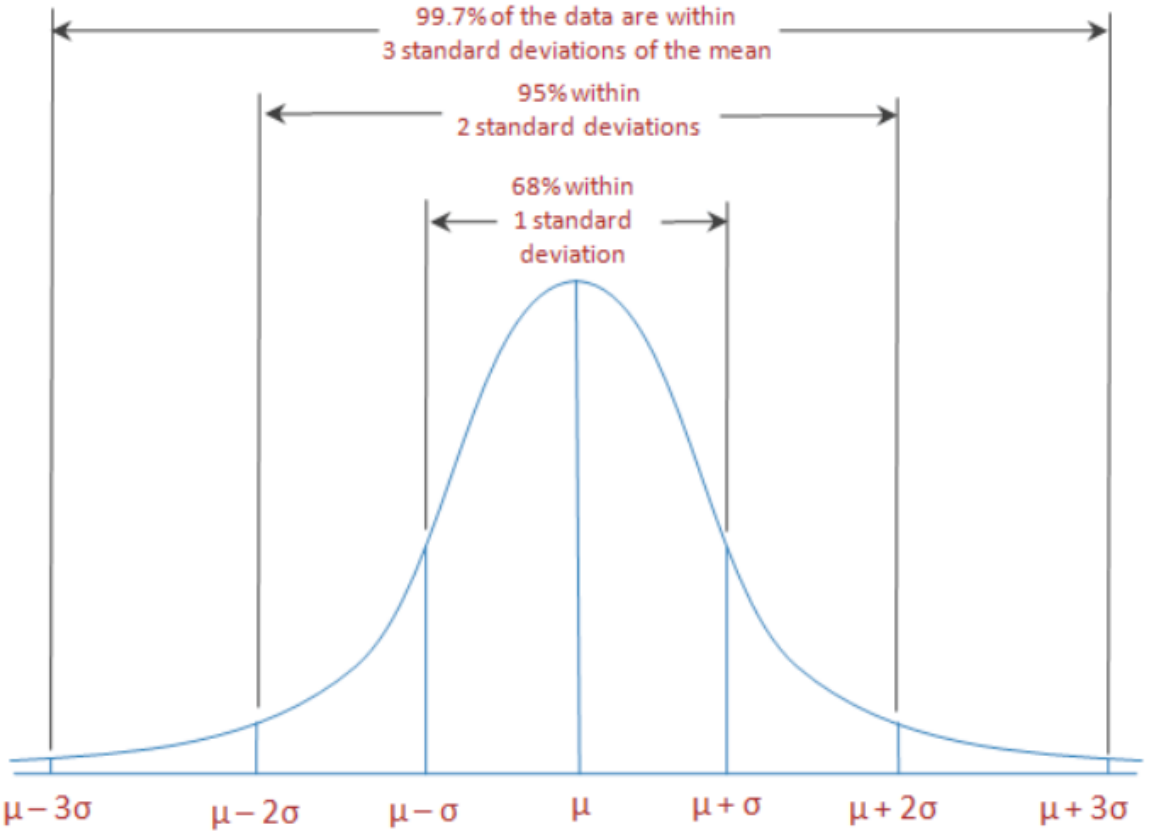
σ = population standard deviation

N = the size of the population

x_i = each value from the population

μ = the population mean

Emprical Rule



Normal distribution & empirical rule (68-95-99.7% rule)

- $\mu \pm \sigma$ includes approximately 68% of the observations
- $\mu \pm 2 \cdot \sigma$ includes approximately 95% of the observations
- $\mu \pm 3 \cdot \sigma$ includes almost all of the observations (99.7% to be more precise)

Mean and Variance

- Mean and variance is a measure of central dispersion.
- Mean is the average of given set of numbers.

- The average of the squared difference from the mean is the variance.

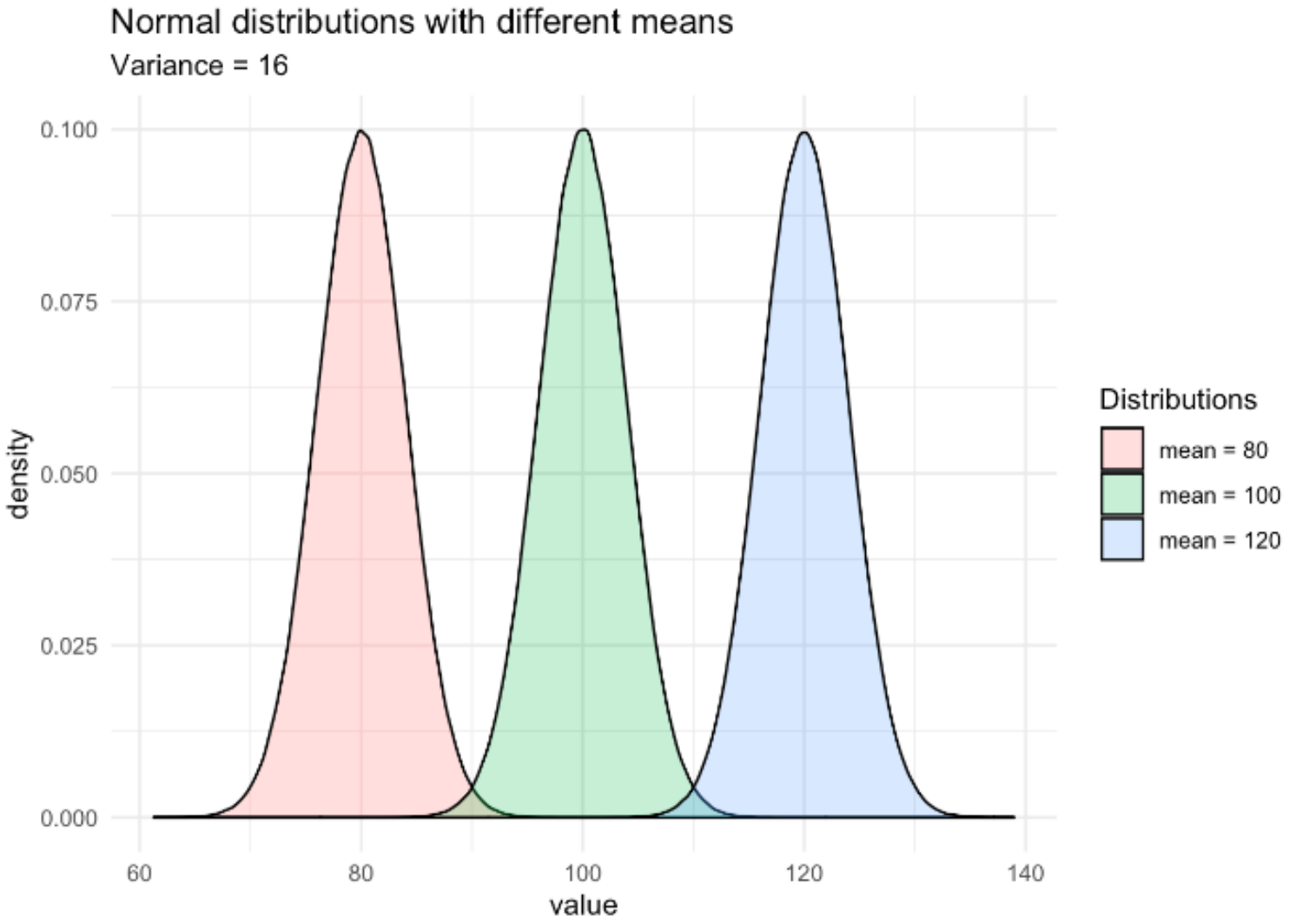
- Central dispersion tells us how the data that we are taking for observation are scattered and distributed.

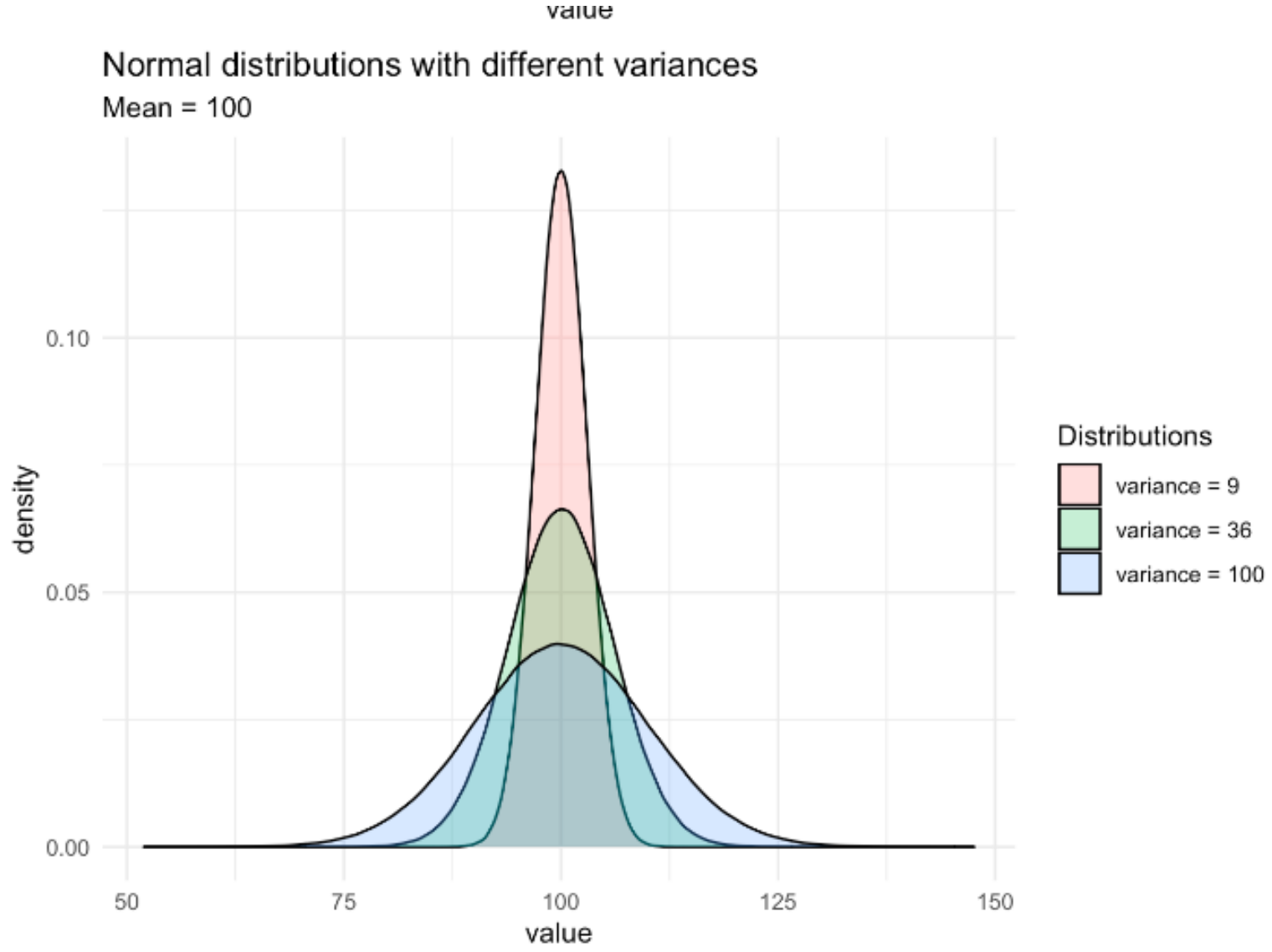
- The variance (σ^2), is defined as the sum of the squared distances of each term in the distribution from the mean (μ), divided by the number of terms in the distribution (N).

Mean and Variance

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

$$\sigma^2 = \frac{\sum X^2}{N} - \mu^2$$


Normal distributions with different means
Variance = 16
0.100
0.075
0.050
0.025
0.000
density
60
80
100
120
140
value
Distributions
mean = 80
mean = 100
mean = 120


value
Normal distributions with different variances
Mean = 100
0.10
0.05
0.00
density
50
75
100
125
150
value
Distributions
variance = 9
variance = 36
variance = 100

Empirical Rule

- Suppose that the scores of CAT1 exam in a course given to all students follows approximately, a normal distribution with mean $\mu=67$ and standard deviation $\sigma=9$.
- It can then be deduced that

- approximately 68% of the scores are between 58 and 76
- approximately 95% of the scores are between 49 and 85, and
- almost all of the scores (99.7%) are between 40 and 94.

Z-Score

- The normal standard distribution is a special case of the normal distribution where the mean is equal to 0 and the variance is equal to 1.
- A normal random variable X can always be transformed to a standard normal random variable Z, a process known as “scaling” or “standardization”
  - by subtracting the mean from the observation, and dividing the result by the standard deviation.

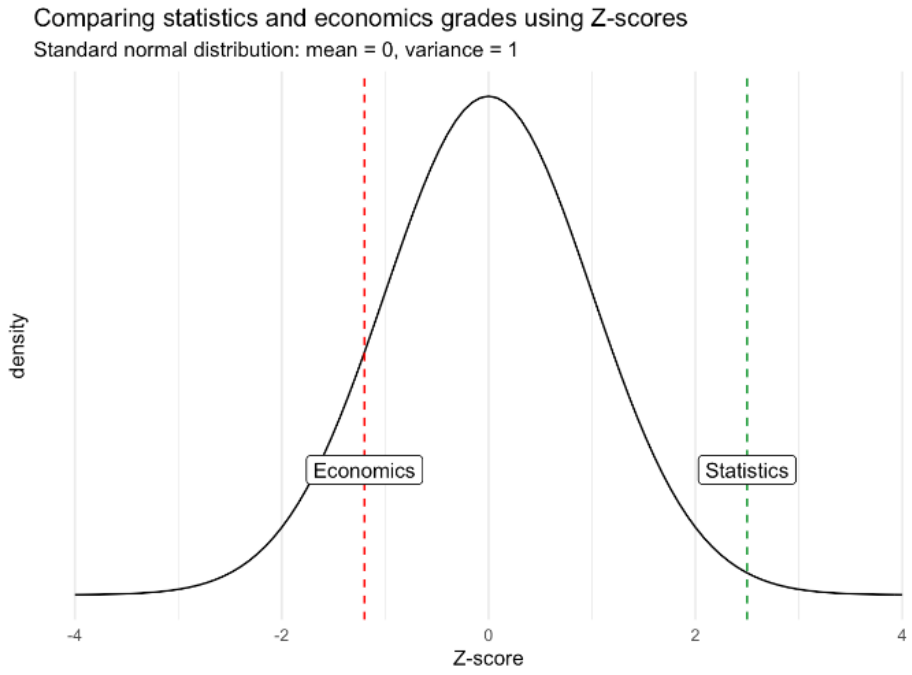
$$Z = \frac{x - \mu}{\sigma}$$

	Statistics	Economics
Mean	40	80
Standard deviation	8	12.5
Student's score	60	65

Z-scores for:

- Statistics: $z_{stat} = \frac{60-40}{8} = 2.5$
- Economics: $z_{econ} = \frac{65-80}{12.5} = -1.2$

On the one hand, the Z-score for the exam in statistics is positive ($z_{stat} = 2.5$) which means that she performed better than average. On the other hand, her score for the exam in economics is negative ($z_{econ} = -1.2$) which means that she performed worse than average.



Although the score in economics is better in absolute terms, the score in statistics is actually relatively better when comparing each score within its own distribution.

- The pnorm function **gives the Cumulative Distribution Function (CDF) of the Normal distribution** in R, which is the probability that the variable X takes a value lower or equal to x.

R Code

Let Z denote a normal random variable with mean 0 and standard deviation 1, find $P(Z > 1)$

```
> pnorm(1,mean=0,sd=1,lower.tail = FALSE)
[1] 0.1586553
```

Let Z denote a normal random variable with mean 0 and standard deviation 1, find $P(-1 \leq Z \leq 1)$.

```
> pnorm(1,lower.tail = TRUE)-pnorm(-1,lower.tail =
 TRUE)
[1] 0.6826895
```

Let Z denote a normal random variable with mean 0 and standard deviation 1, find $P(0 \leq Z \leq 1.37)$.

```
> pnorm(0,lower.tail = FALSE)-pnorm(1.37,lower.tail
 = FALSE)
[1] 0.4146565
```

$P(70 \leq X \leq 80)$ where $X \sim \mathcal{N}(\mu = 67, \sigma^2 = 9^2)$

```
> pnorm(70,mean=67,sd=9,lower.tail = FALSE)-pnorm(80,
mean=67,sd=9,lower.tail = FALSE)
[1] 0.2951343
```

Computing Skewness

- The **moment coefficient of skewness** of a data set is
- skewness: $g_1 = m_3 / m_2^{3/2}$
- where
- $m_3 = \sum(x-\bar{x})^3 / n$ and $m_2 = \sum(x-\bar{x})^2 / n$
- $\bar{x}$ is the mean and n is the sample size, as usual. m_3 is called the **third moment** of the data set. m_2 is the **variance**, the square of the standard deviation.
- If you have the whole population, then g_1 above is the measure of skewness. But **if you have just a sample**, you need the **sample skewness**:

$$\text{sample skewness: } G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

Are my data normal?

- **Example**: Let's consider the example of the college men's heights, and compute the skewness.

Height (inches)	Class Mark, x	Frequency, f
59.5–62.5	61	5
62.5–65.5	64	18
65.5–68.5	67	42
68.5–71.5	70	27
71.5–74.5	73	8

- Begin with the sample size and sample mean.

- $n = 5+18+42+27+8 = 100$
- $\bar{x} = (61\times5 + 64\times18 + 67\times42 + 70\times27 + 73\times8) \div 100$
- $\bar{x} = 9305 + 1152 + 2814 + 1890 + 584) \div 100$
- $\bar{x} = 6745 \div 100 = 67.45$

Class Mark, x	Frequency, f	xf	$(x-\bar{x})$	$(x-\bar{x})^2 f$	$(x-\bar{x})^3 f$
61	5	305	-6.45	208.01	-1341.68
64	18	1152	-3.45	214.25	-739.15
67	42	2814	-0.45	8.51	-3.83
70	27	1890	2.55	175.57	447.70
73	8	584	5.55	246.42	1367.63
	Σ	6745	n/a	852.75	−269.33
	$\bar{x}, m_2, m_3$	67.45	n/a	8.5275	−2.6933

- Finally, the population skewness is
- $g_1 = m_3 / m_2^{3/2} = -2.6933 / 8.5275^{3/2} = -0.1082$

sample skewness:

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1 = [\sqrt{100 \times 99} / 98] [-2.6933 / 8.5275^{3/2}] = -0.1098$$

Interpreting

- If skewness is positive, the data are positively skewed or skewed right, meaning that the right tail of the distribution is longer than the left. If skewness is negative, the data are negatively skewed or skewed left, meaning that the left tail is longer.
- If skewness = 0, the data are perfectly symmetrical. But a skewness of exactly zero is quite unlikely for real-world data.
- If skewness is less than −1 or greater than +1, the distribution can be called **highly skewed**.
- If skewness is between −1 and −½ or between +½ and +1, the distribution can be called **moderately skewed**.
- If skewness is between −½ and +½, the distribution can be called **approximately symmetric**.
- With a skewness of −0.1098, the sample data for student heights are approximately symmetric.

Reference

- https://statsandr.com/blog/do-my-data-follow-a-normal-distribution-a-note-on-the-most-widely-used-distribution-and-how-to-test-for-normality-in-r/
- https://statisticsbyjim.com/probability/empirical-rule/