Statistical Literacy — MINT

Lecture 5: Continuous random variables and the Normal distribution

Rémi Viné

The Graduate Institute | Geneva

October 23rd, 2023

Outline

Housekeeping

- Problem set 5 is now available
- Next week's class will be in A1B due to Geneva Peace Week

Random Variables

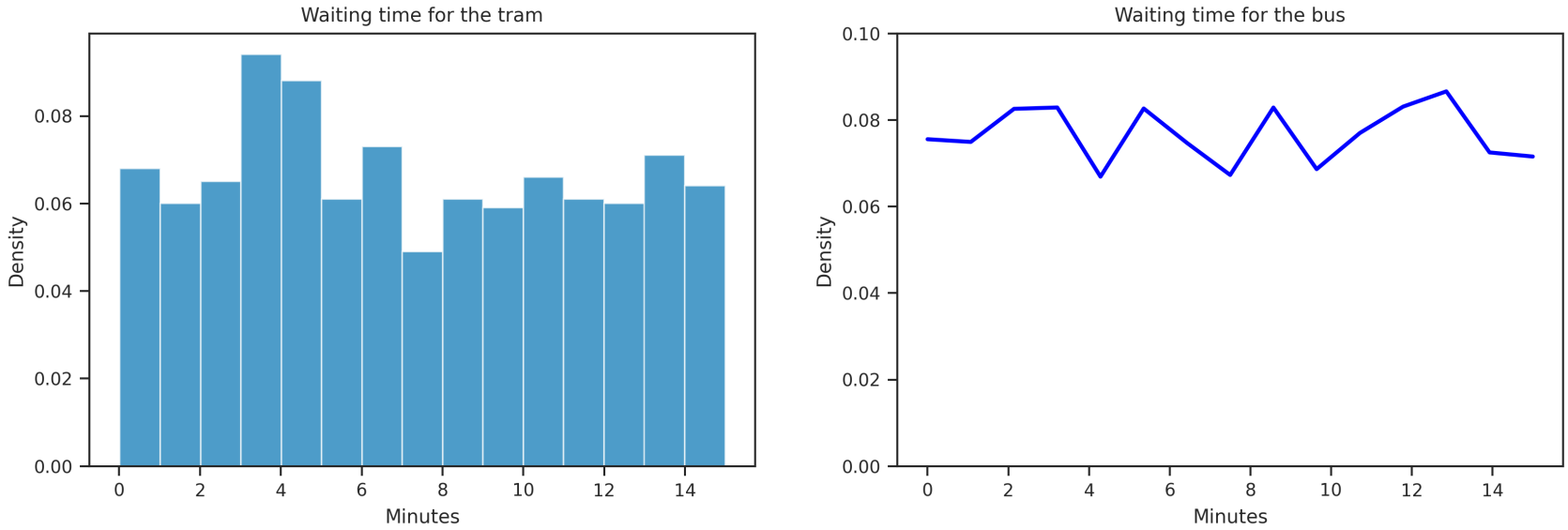
Discrete	Rolling a die
	Number of casualties in a bombing
	Passengers in a flight
	Number of votes in favor of a resolution
Continuous	Noise in a flight
	Waiting time for the bus
	Radioactivity levels next to a uranium mine
	Exchange rate CHF-€

Recall of the continuous case

- Unlike discrete variables, continuous variables are not countable
- They can take an infinite number of values
  - There is an infinity of $P(X = x_i)$ so that each cannot be granted a probability
- For continuous random variables, **intervals** are of interest
  - $P(X < x_i)$
  - $P(X > x_i)$
  - $P(x_i < X < x_j)$

Recall of the continuous case

Example



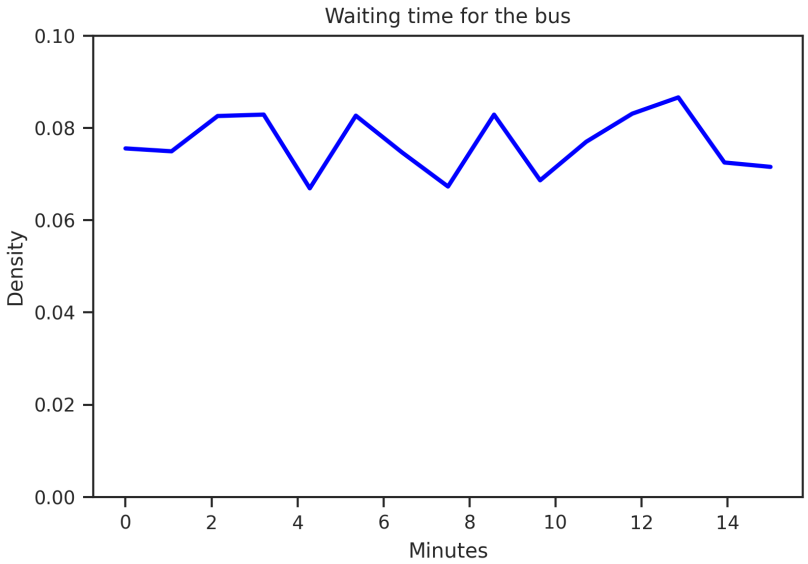
(a) Histogram

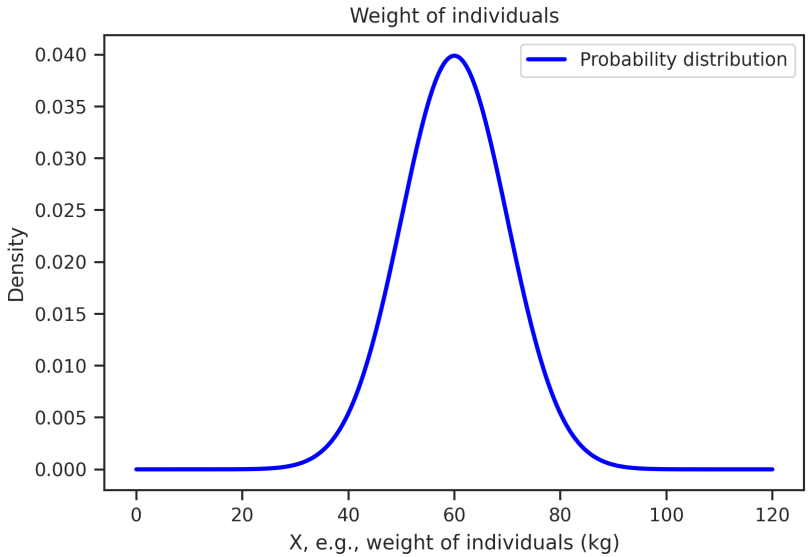
(b) Density

- Here, one cannot tell the probability to wait π minutes, 1 minute, or $\sqrt{2}$ minutes (although these are possible)
- But one could tell *(using a software or some math)*
  - $P(X < 8)$, $P(X > 5)$, $P(2 < X < 9)$...

The density plot

- Often continuous distributions are represented using **density plots**
  - This is a line (curve) whose area under it is 1
    - Remember Kolmogorov?

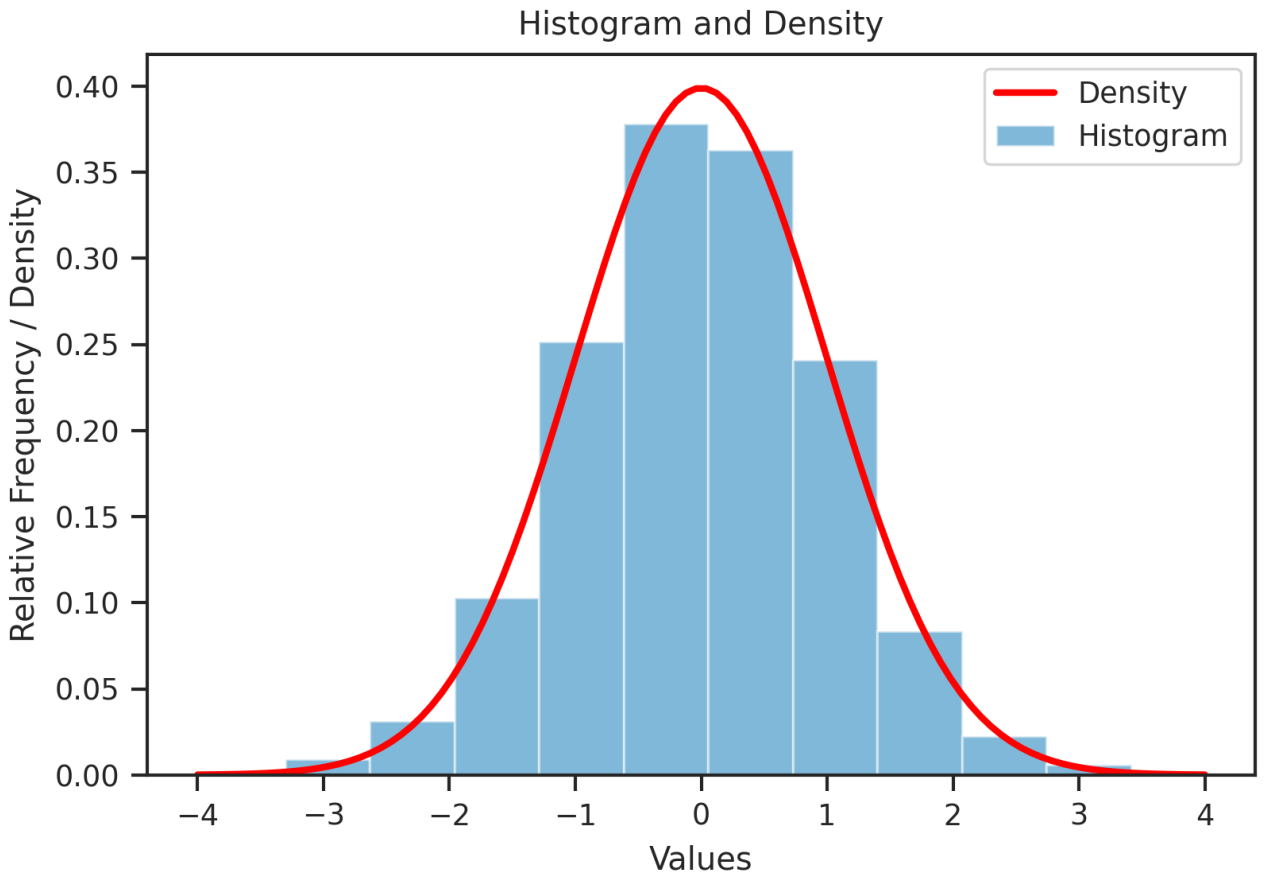




- The probability of observing a situation that lies within a given interval is the area of this interval under the curve

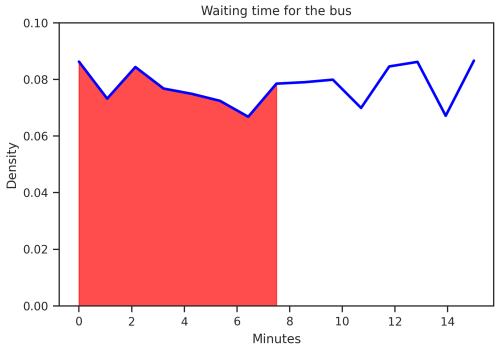
The density plot and the histogram

- Imagine cutting pieces of the histogram and putting them under the curve, this should overlap (by construction) - or almost overlap...

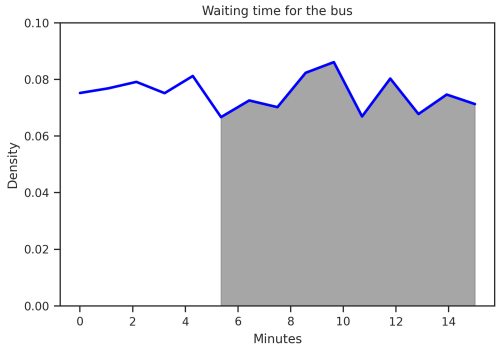


Recall of the continuous case

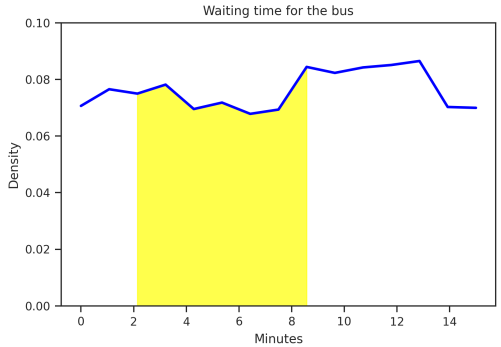
Example



(a) $P(X < 7.8)$



(b) $P(X > 5.5)$

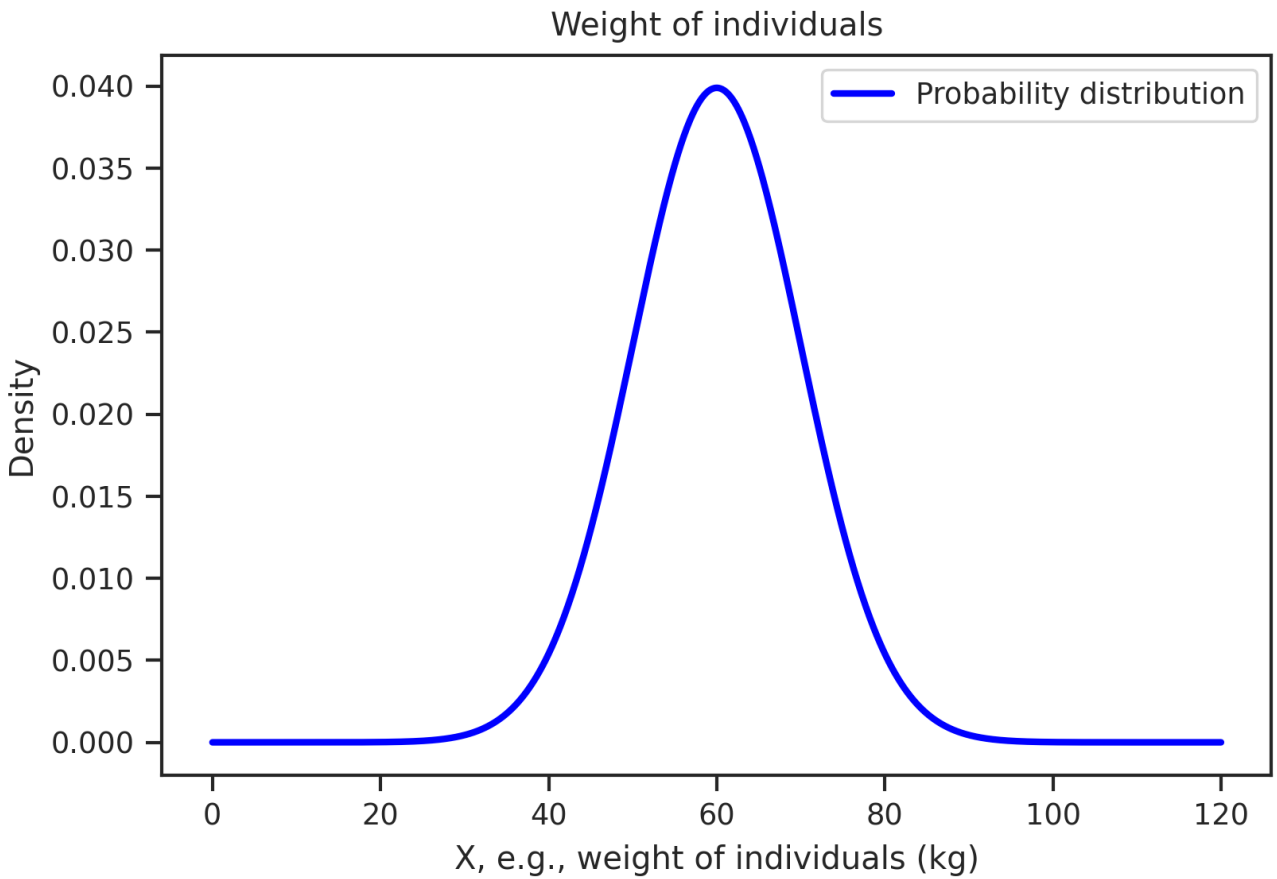


(c) $P(2.1 < X < 8.5)$

- **A lot of statistics require to know the share (of the total area under the curve) that the shaded area of interest represents**
  - Sometimes the bound(s) is(are) defined (*e.g.*, $P(X > 5.5) = ?$)
  - Sometimes the share is defined (*e.g.*, $P(X > ?) = 0.31$)

The Normal Distribution

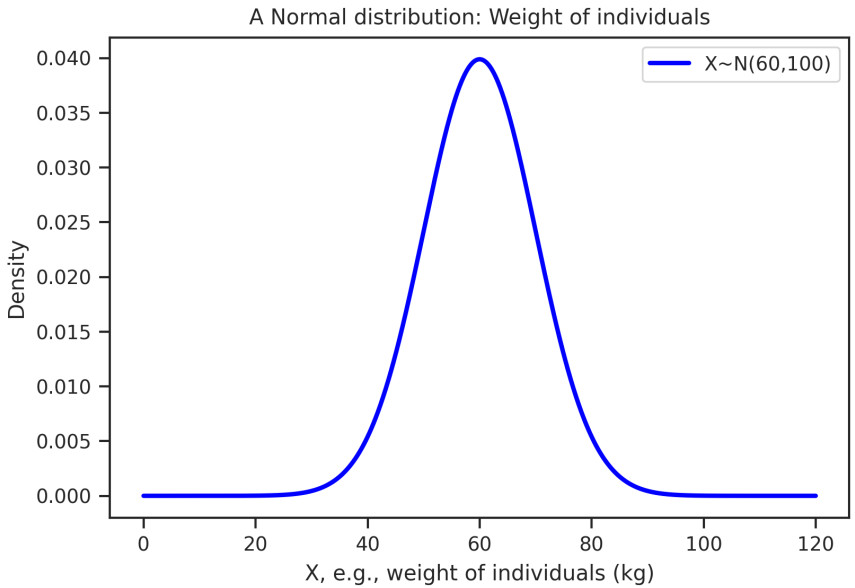
- This is a bell-shaped distribution
  - $Average = Median = Mode$
  - The empirical rule applies



The Normal Distribution - formally

- The density function depends on 2 parameters: μ and σ:

$$If\ X \sim N(\mu, \sigma^2),\ then\ f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$



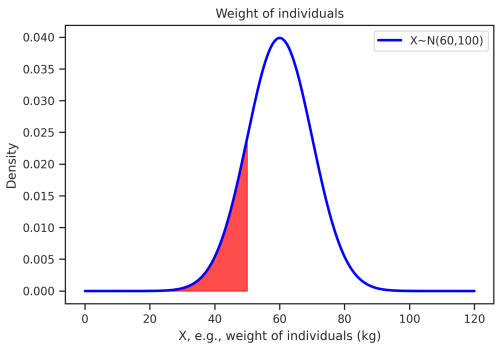
- A random variable that follows a Normal distribution of mean 60 and standard deviation 10 is denoted as:

$$X \sim N(60, 100)$$

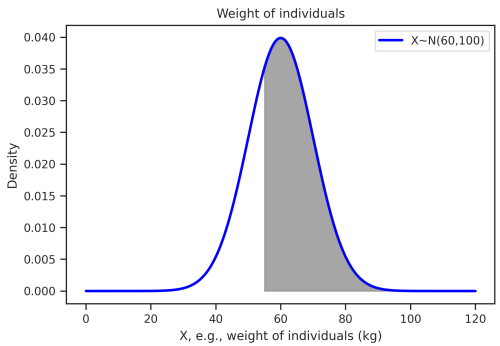
The Normal Distribution - what for?

- The idea is to compute the area under the curve that is of interest

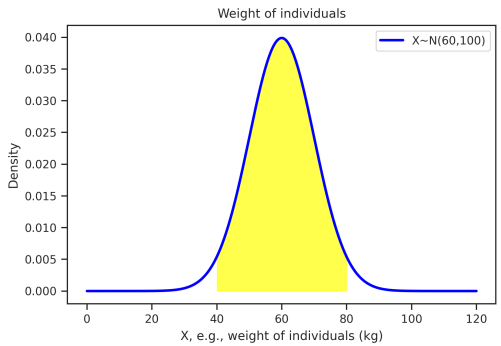
$$P(x_i < X < x_j) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{x_i}^{x_j} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$



(a) $P(X < 50)$



(b) $P(X > 55)$



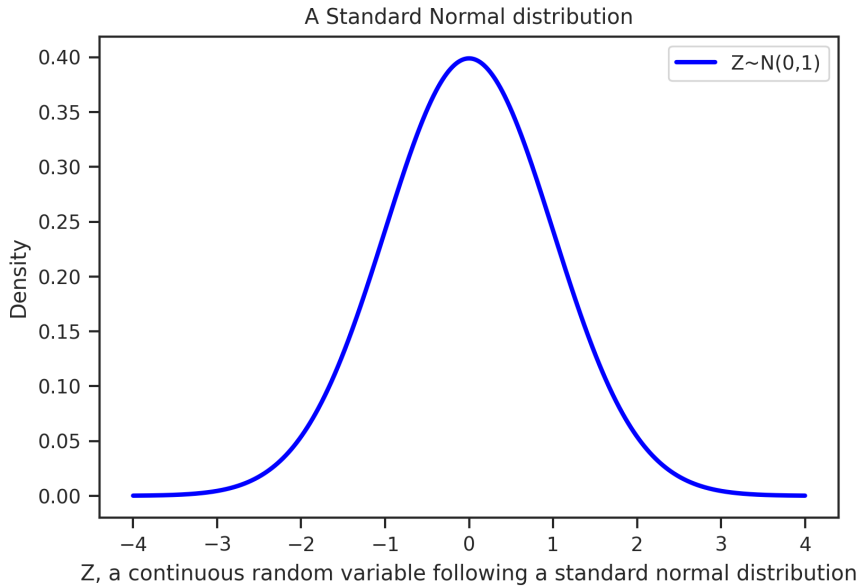
(c) $P(40 < X < 80)$

- **Challenge**: can you find what is $P(40 < X < 80)$?

The **Standard** Normal Distribution

- This is a special Normal Distribution, the random variable is usually denoted Z
  - The average is 0: $\mu = 0$
  - The standard deviation is 1: $\sigma = 1$ *(hence, the variance is 1 as well)*

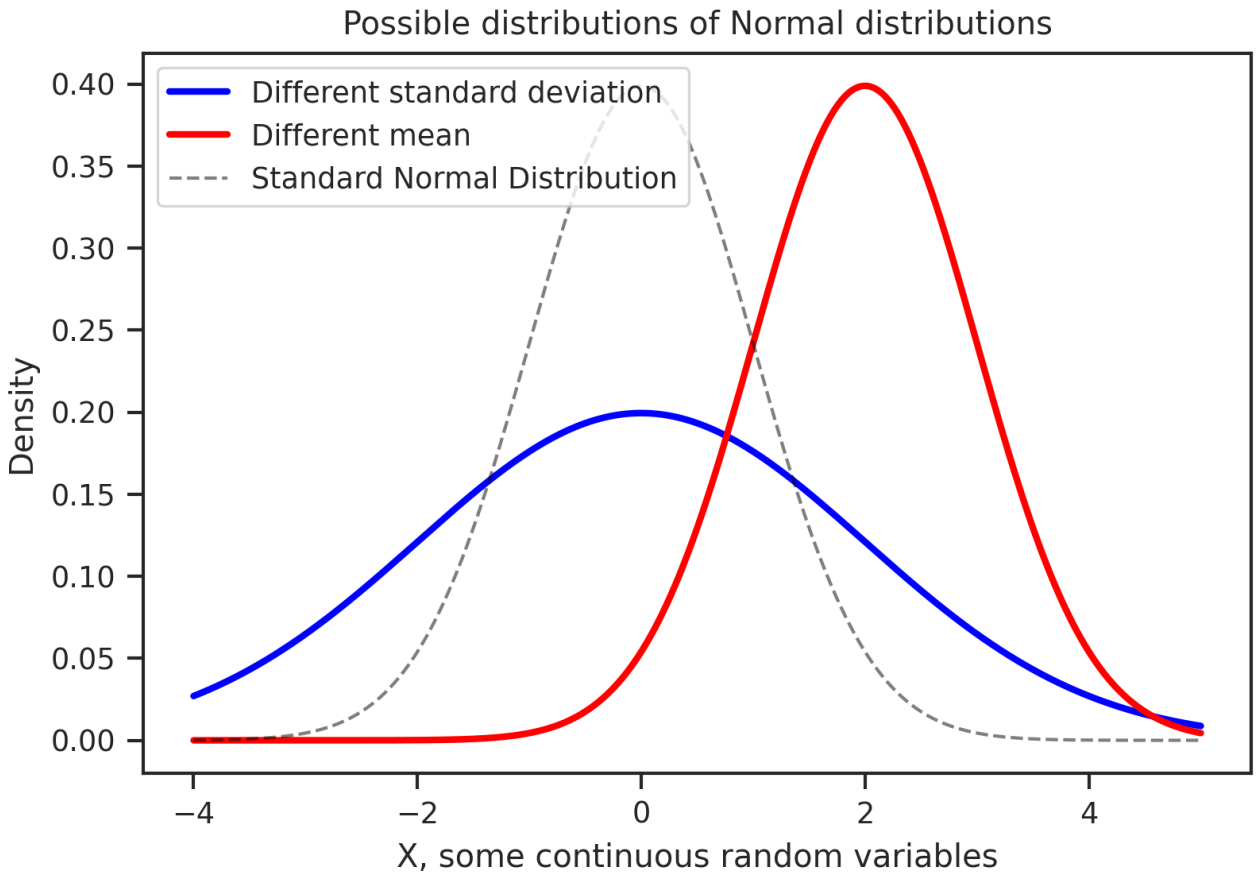
$$If\ Z \sim N(0,1),\ then\ f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$



- Units of the Z distribution are expressed in **standard deviations**
- For $X \sim N(\mu, \sigma^2)$, values of X are expressed in the units of the variable

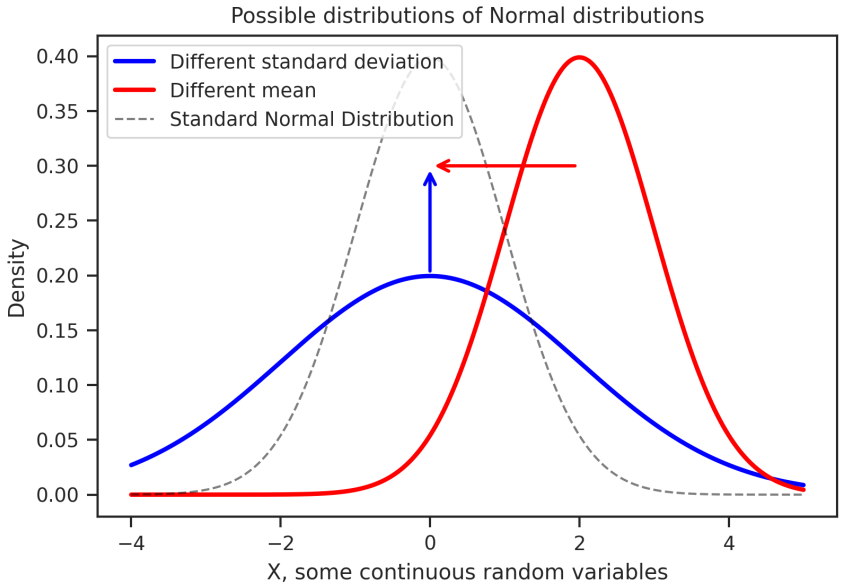
All Normal Distributions are not Standard

- For non-Standard Normal Distributions:
  - The mean can differ from 0
  - The standard deviation can differ from 1



Standardizing

- Mixing apples and oranges on purpose
  - All Normal Distributions, once standardized, are expressed in the same magnitude
  - Standardized units of 1kg, 1m, 1hertz, 1becquerel, 1volt, *etc.* are all expressed in units of standard deviations



- From $X \sim N(\mu, \sigma^2)$ to $Z \sim N(0, 1)$:

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

- *From the z-score to the X observation:*

$$x = \mu + \sigma \times z$$

Standardizing

Examples

1. If the weight of students in IHEID is Normally distributed with an average of 60kg and a standard deviation of 20kg, what is the standardized weight of a person of 72kg?
2. If the CHF-€ exchange rate follows a Normal distribution, $X \sim N(1.05, 0.02)$, what is the standardized rate of an exchange equal to 0.99?
3. Assume the standardized distance to work in Geneva (for a given individual) is -2.
   - 3.1 What does it mean?
   - 3.2 If the average distance to work is 15km and the variance is $16km^2$, what is the distance to work in Geneva for the individual (whose standardized distance is -2)?

Standardizing

Examples

1. If the weight of students in IHEID is Normally distributed with an average of 60kg and a standard deviation of 20kg, what is the standardized weight of a person of 72kg?
   - $z = \frac{72-60}{20} = 0.6$
2. If the CHF-€ exchange rate follows a Normal distribution, $X \sim N(1.05, 0.02)$, what is the standardized rate of an exchange equal to 0.99?
   - $z = \frac{0.99-1.05}{\sqrt{0.02}} \approx -0.424$

Standardizing

Examples

3. Assume the standardized distance to work in Geneva (for a given individual) is -2.

   3.1 What does it mean?

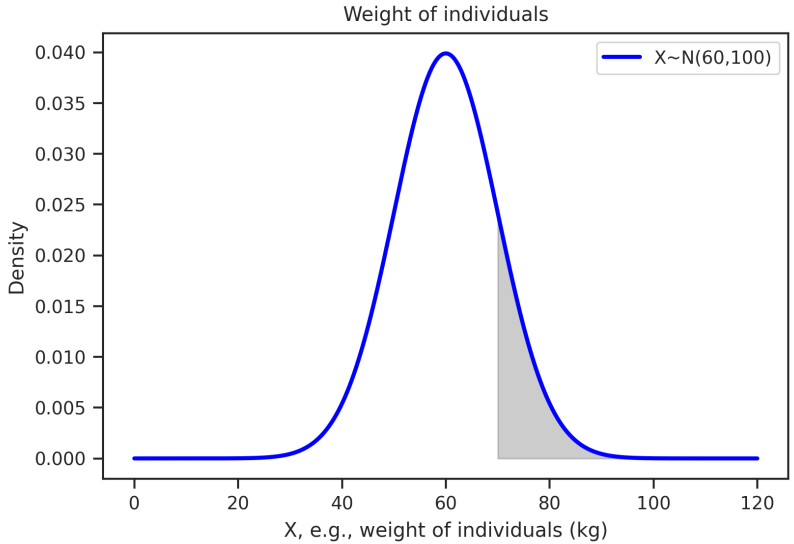
   - Geneva's distance to work for this individual is 2 standard deviations lower than for the average of other individuals
   - Using the Empirical rule, s/he belongs to the 2.5% individuals with lower distances to work *(this is not real data!)*

   3.2 If the average distance to work is 15km and the variance is $16km^2$, what is the distance to work for him or her?

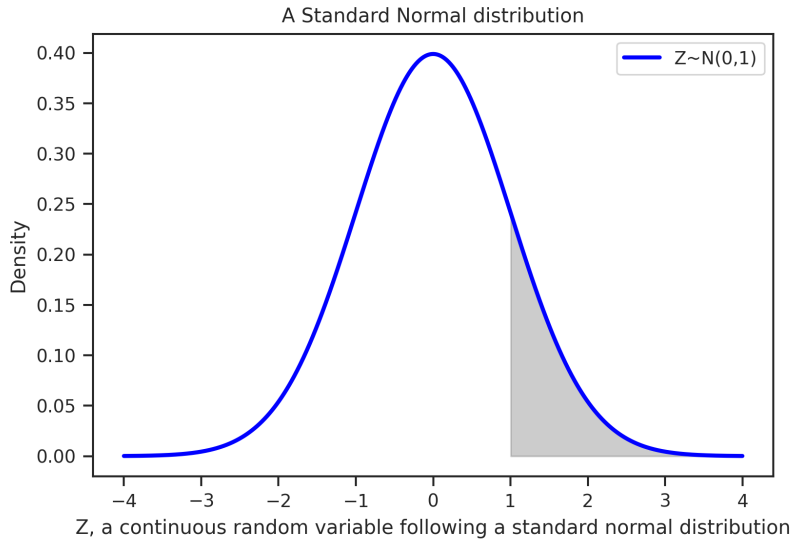
   - $x = \mu + \sigma z = 15 + \sqrt{16} \times (-2) = 7$
   - Hence the distance to work for this person would be 7km

Find probabilities

What is the probability to select (randomly) a person weighting more than 70kg?



(a) $X \sim N(60, 100)$



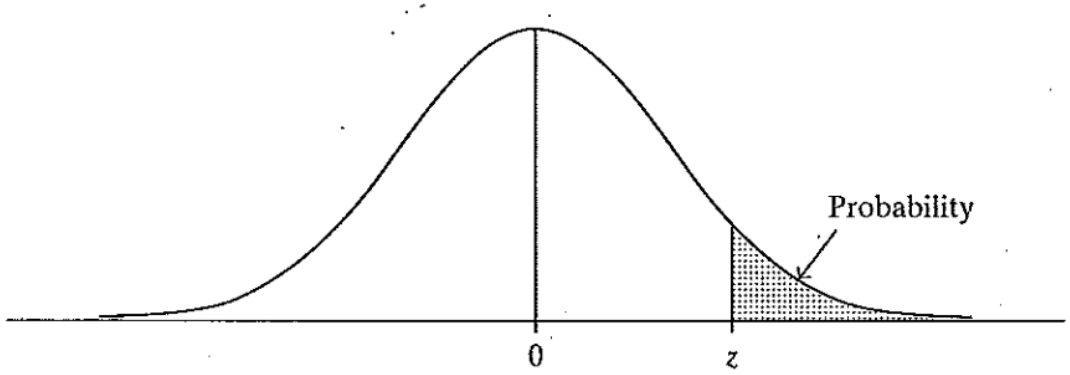
(b) $Z \sim N(0, 1)$

- $P(X > 70) = P(Z > \frac{70-60}{10}) = P(Z > 1)$
- From this, one can use a table (or use a statistical software)

Find probabilities

What is the probability to select (randomly) a person weighting more than 70kg?

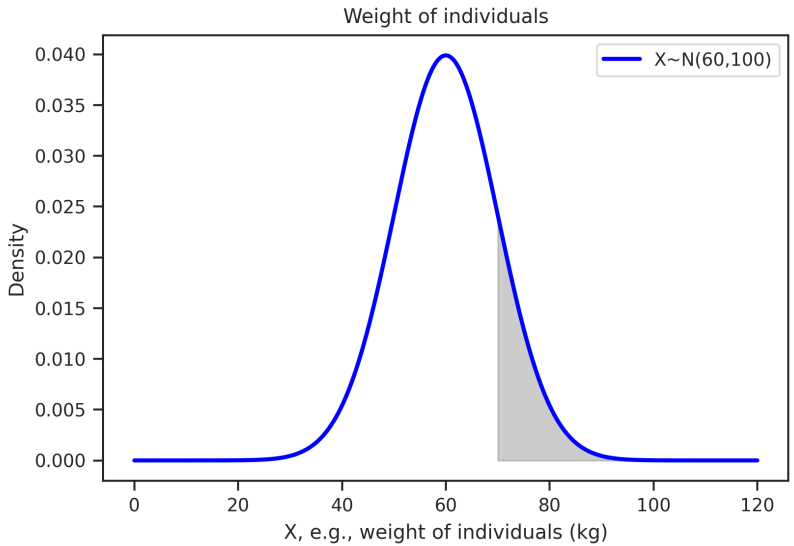
TABLE A: Normal curve tail probabilities. Standard normal probability in right-hand tail (for negative values of z, probabilities are found by symmetry)



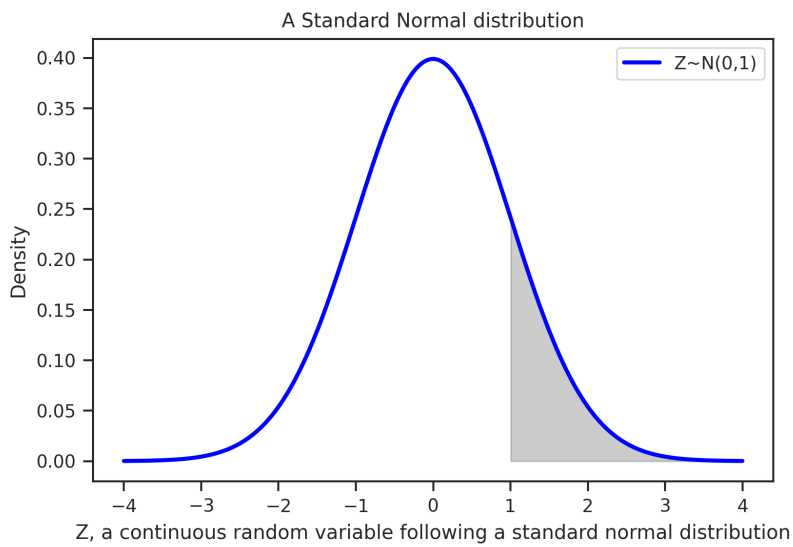
	Second Decimal Place of z									
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170

Find probabilities

What is the probability to select (randomly) a person weighting more than 70kg?



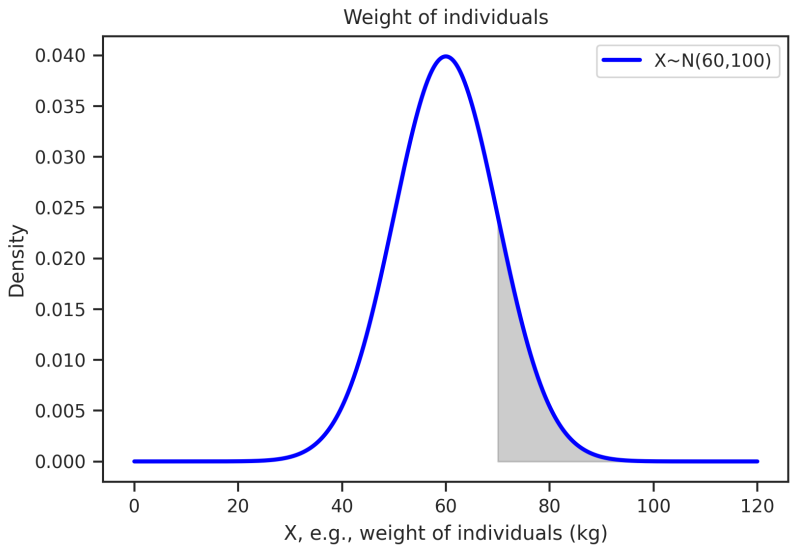
(a) $X \sim N(60, 100)$



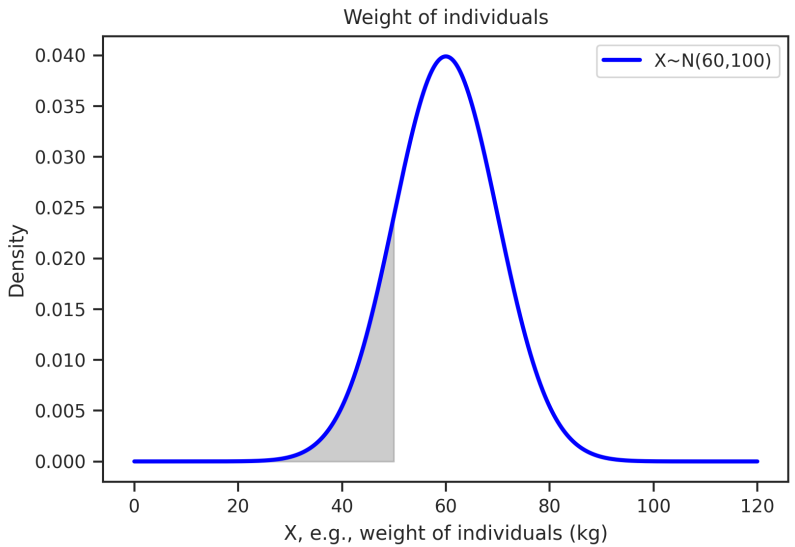
(b) $Z \sim N(0, 1)$

- Using a standard normal distribution table:
  - $P(X > 70) = P(Z > 1) = 15.87\%$
  - Almost 16% of individuals weight more than 70kg

Find probabilities: 15.87% of people weighting less than 50kg



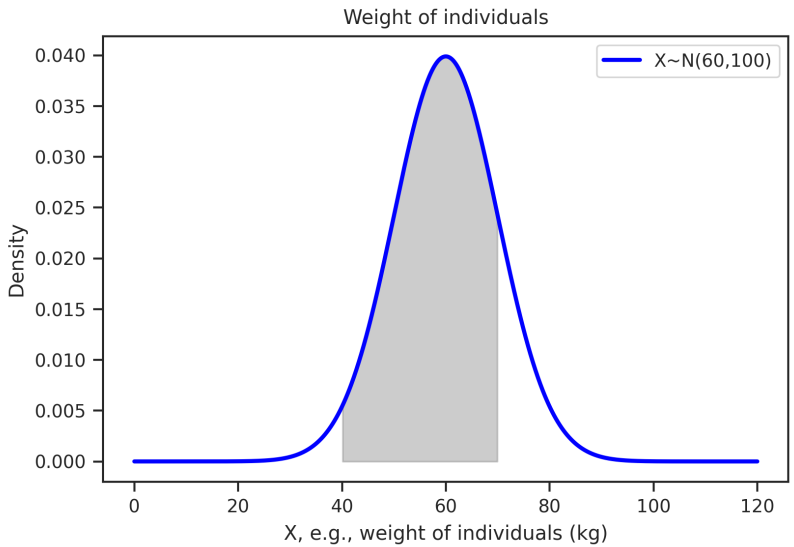
(a) $P(X > 70)$



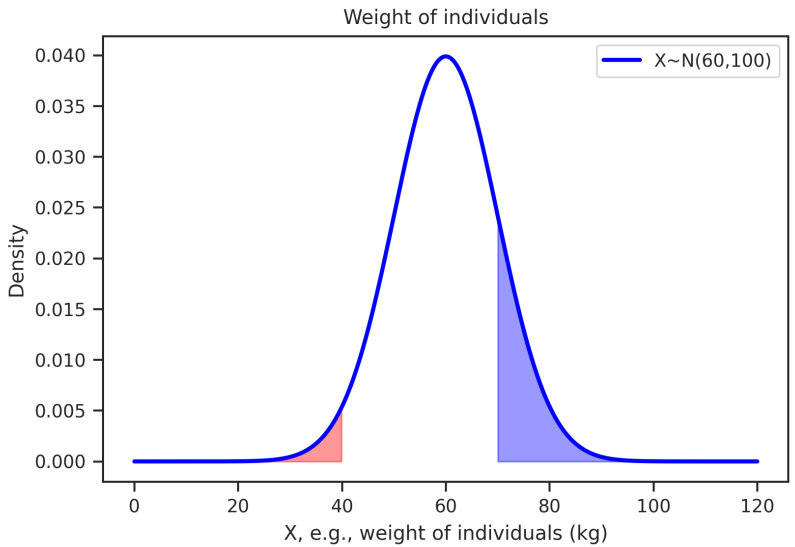
(b) $P(X < 50)$

- By symmetry, these two areas are similar
- Hence, using the table, the *z-score* of 1 can be used, as $P(Z < -1) = P(Z > 1)$

Find probabilities: share of people weigh 40 to 70kg



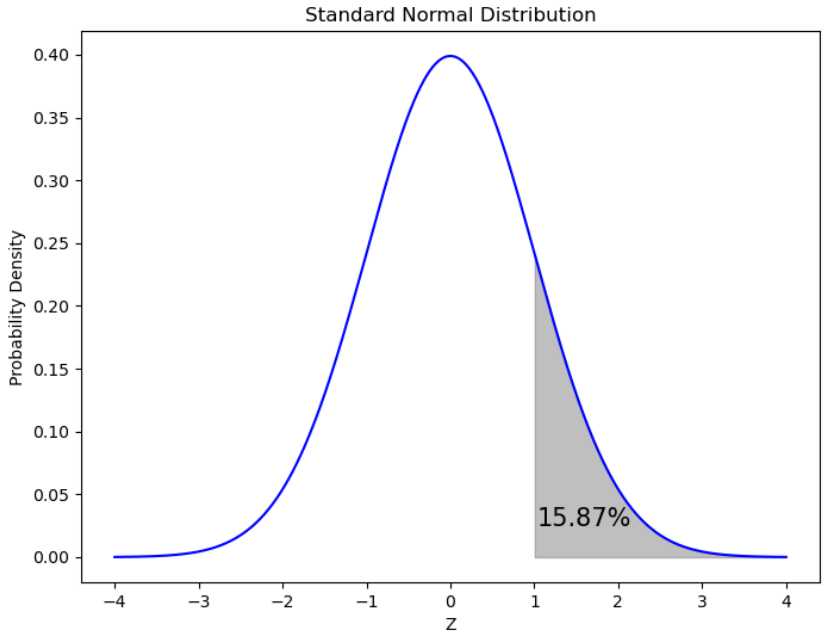
(a) The area of interest



(b) The complement areas

- $P(40 < X < 70) = 1 - (P(X < 40) + P(X > 70)) = 1 - P(Z < -2) - P(Z > 1) = 1 - P(Z > 2) - P(Z > 1) = 1 - 0.0228 - 0.1587 = 0.8185$
- Hence, 81.85% of people weight between 40kg and 70kg

Find *z-score* from a probability



- $P(Z > z) = 0.1587$: regarding where on the table is 0.1587, it corresponds to $z = 1$
- Easy to *destandardize*: $x = \mu + \sigma \times z$
  - In the weight example: $x = 60 + 10 \times 1 = 70$ (kg)

Next session

- Next session is on the sampling distribution, perhaps the most important (and conceptual) tool we cover in this class...