SM-1402 Basic Statistics

Chapter 3: The normal distribution

Dr. Haziq Jamil
Mathematical Sciences, Faculty of Science, UBD
https://haziqj.ml

Semester II 2022/23

Learning outcomes

- Characterize the normal distribution by its two parameters, and understand the effect of changing these two values.
- Understand the role of the normal distribution function and its relation to areas and probabilities.
- Use the standard normal table to calculate normal probabilities and also use it in reverse to find quantiles.
- Be able to simply evaluate the assumption of normality for data.
- Caculate the point estimate for the mean as well as the 95% confidence interval for it.

Required reading

- Madsen (2016) Chapter 4 (4.1–4.8 only).

Motivation

You buy a 500 g bag of coffee. You (may or) may not be surprised if the weight is not exactly 500 g.

Motivation

You buy a 500 g bag of coffee. You (may or) may not be surprised if the weight is not exactly 500 g.

- If you purchase multiple bags of coffee, expect:
  - the *average* weight to be close to 500 g.
  - the *spread* of the weights (data) is not too large.

    E.g. 501 502 498 500 499 503 500 500 497 500

Motivation

You buy a 500 g bag of coffee. You (may or) may not be surprised if the weight is not exactly 500 g.

- If you purchase multiple bags of coffee, expect:
  - the *average* weight to be close to 500 g.
  - the *spread* of the weights (data) is not too large.

    E.g. 501 502 498 500 499 503 500 500 497 500
- This variation can be modelled by a *statistical distribution*–in particular, the normal distribution.
- By doing so, we can analyse and make deductions easily.

The normal distribution

- Data X that follows a normal distribution will show a *distribution curve* that is symmetrical and "bell-shaped".
- Many naturally occurring phenomena can be modelled as following a normal distribution.
- It is one of the most important distributions in all of statistics!

The normal distribution

- Data X that follows a normal distribution will show a *distribution curve* that is symmetrical and "bell-shaped".
- Many naturally occurring phenomena can be modelled as following a normal distribution.
- It is one of the most important distributions in all of statistics!

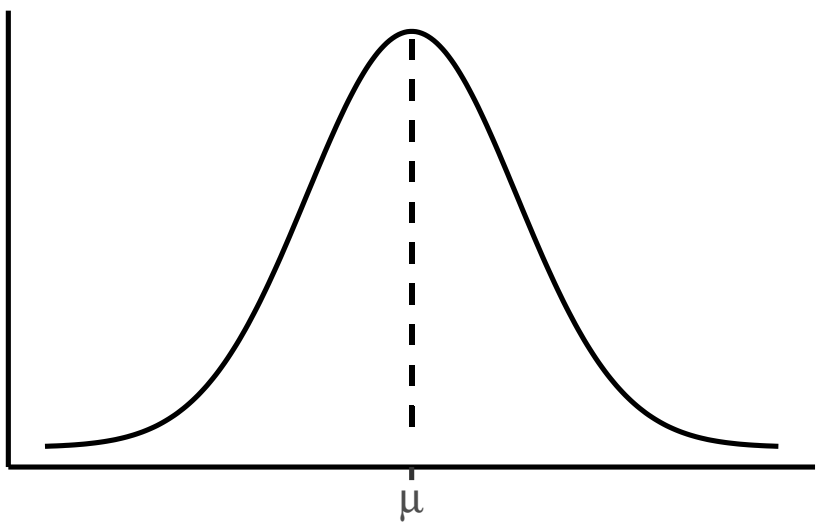
Definition 1

The normal distribution is completely defined by two parameters:

i. Mean μ (miu)–the "center"

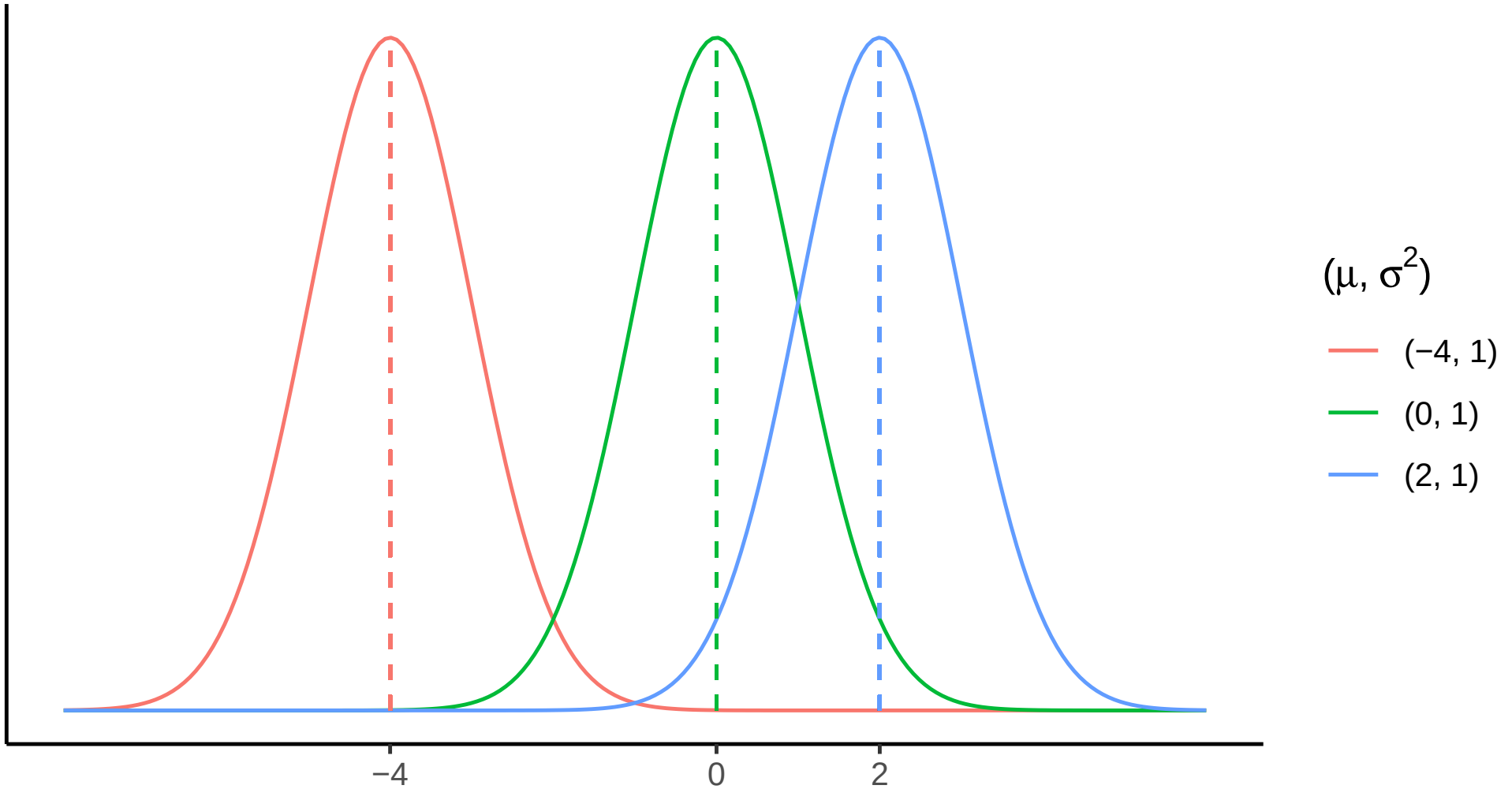
ii. SD σ (sigma)–the "spread"

We write $X \sim \mathrm{N}(\mu, \sigma^2)$ if the distribution of X follows the normal distribution curve.



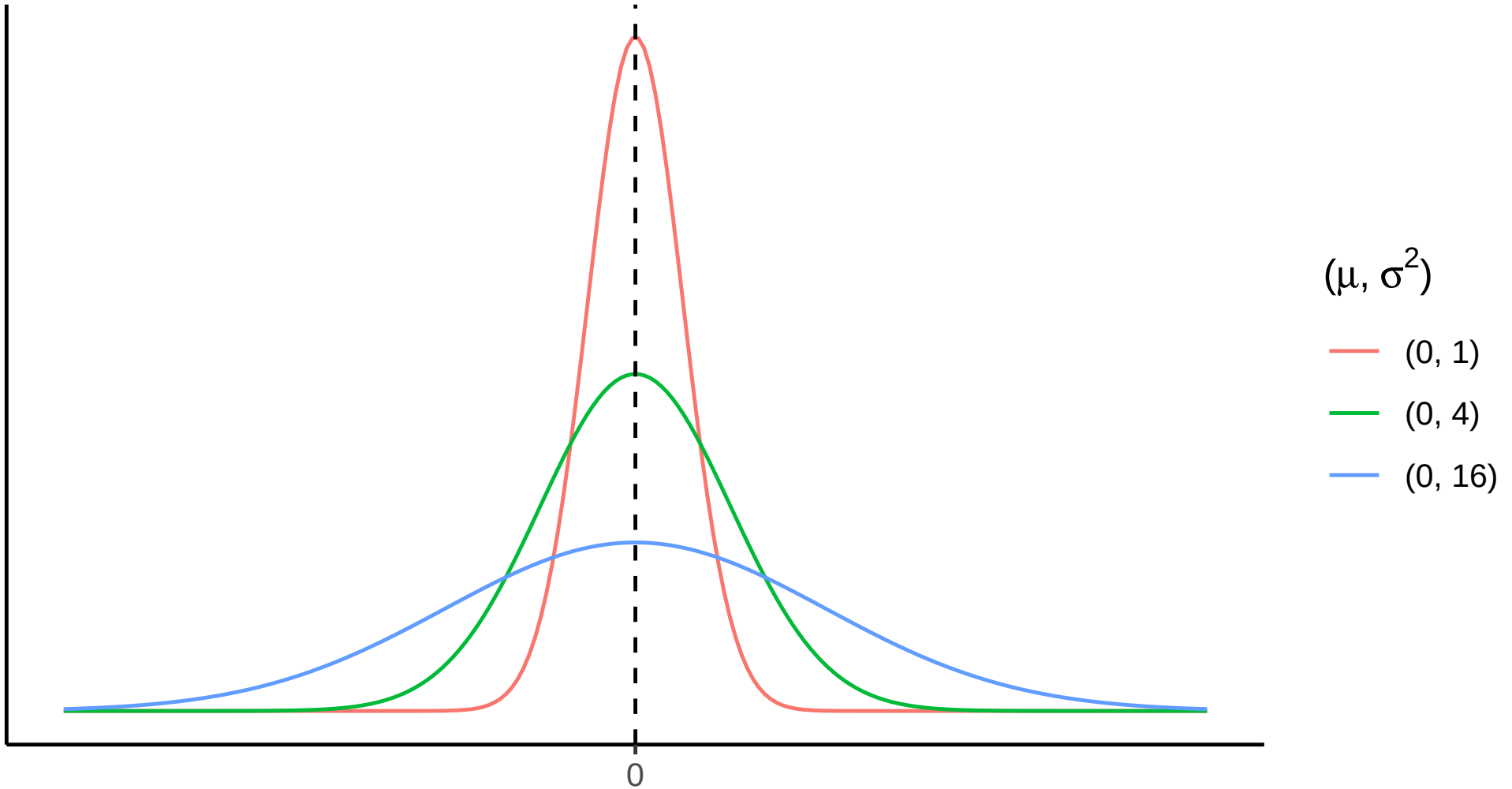
Location parameter

The μ parameter is also called the "location" parameter, since it determines where the bell curve is placed.



Scale parameter

The σ parameter is also called the "scale" parameter, since changing this value scales (spreads) the bell curve accordingly.

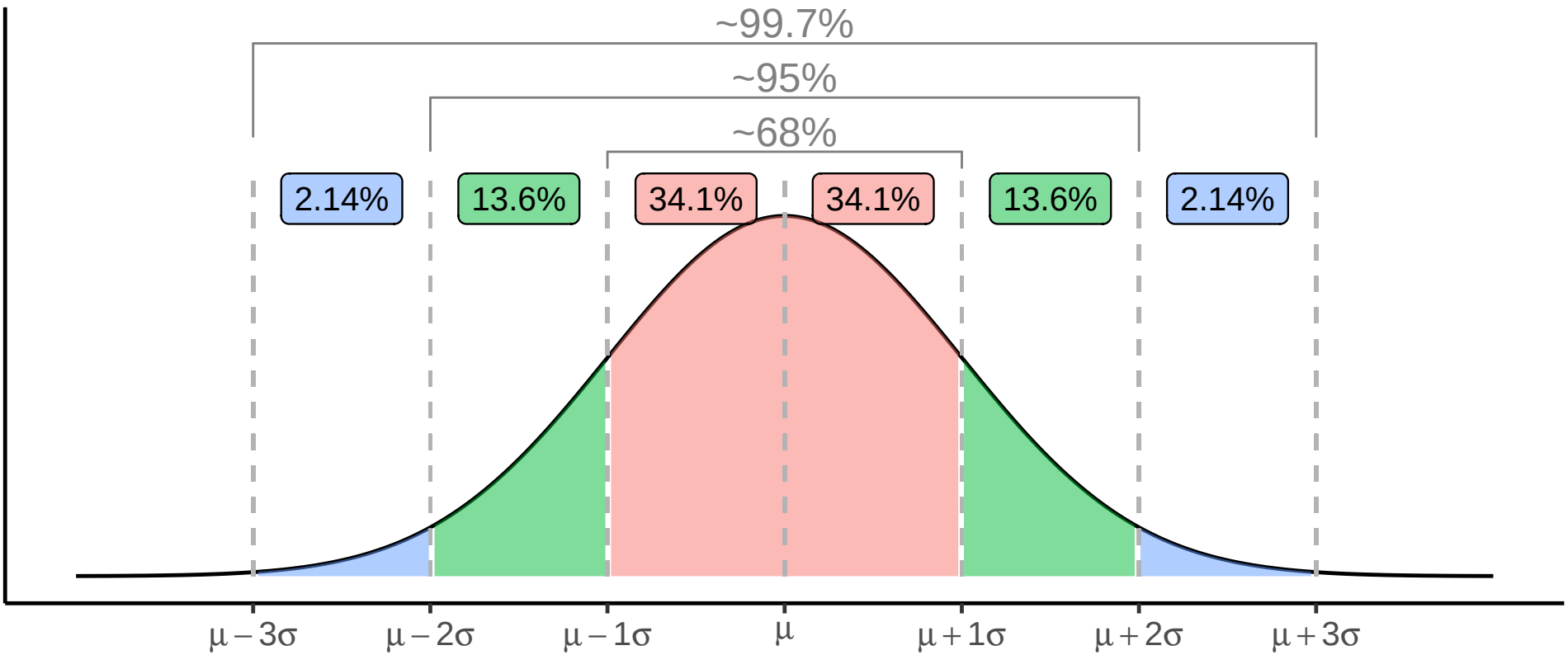


68–95–99.7 Rule

For data following the normal distribution,

- ~68% of data values are within $\mu \pm 1$ SD
- ~95% of data values are within $\mu \pm 2$ SD
- ~99.7% of data values are within $\mu \pm 3$ SD

These percentages are unique only to the normal distribution!



The standard normal distribution

While μ and σ can virtually take any value, there is a pair of values that is especially important: $\mu = 0$ and $\sigma = 1$.

Definition 2 (Standard normal)

Data Z is said to follow the *standard normal* distribution if $Z \sim \mathrm{N}(0, 1)$.

The standard normal distribution

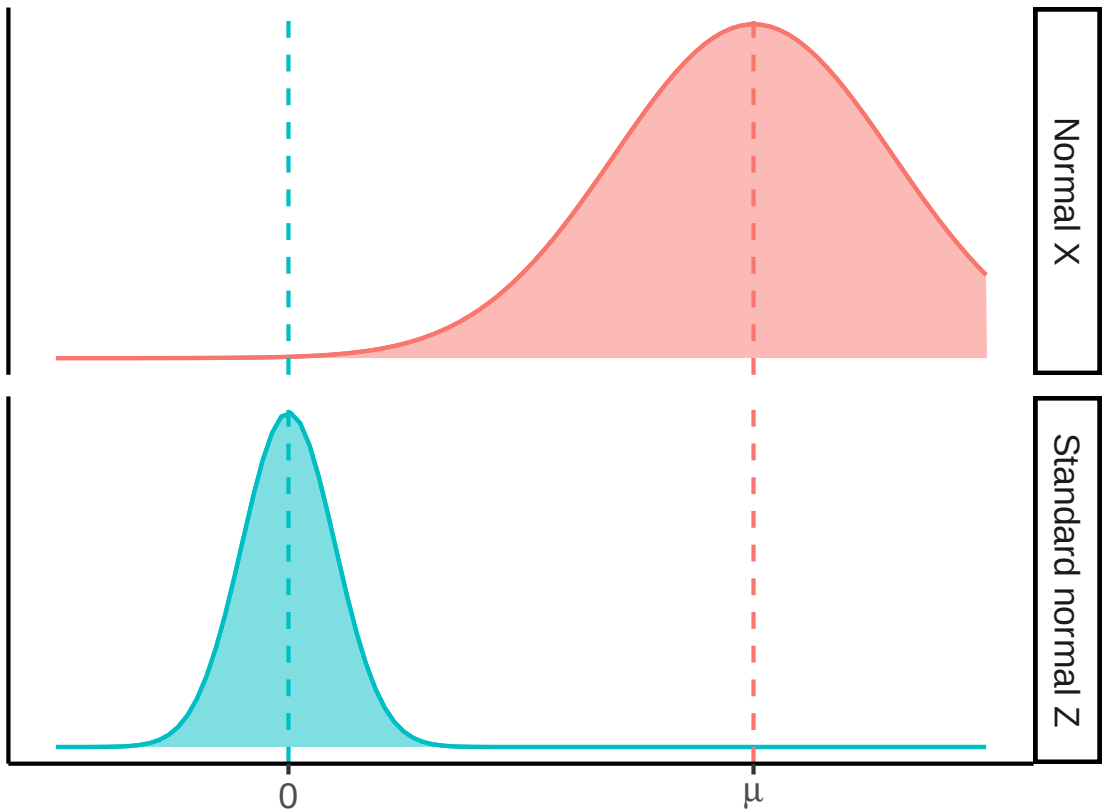
While μ and σ can virtually take any value, there is a pair of values that is especially important: $\mu = 0$ and $\sigma = 1$.

Definition 2 (Standard normal)

Data Z is said to follow the *standard normal* distribution if $Z \sim \mathrm{N}(0, 1)$.

Any normal variable X can be transformed to a standard normal via

In essence, we only ever deal with one normal dist!



Density and distribution function

Areas under the normal curve–representing *probabilities*–are of interest.

Definition 3

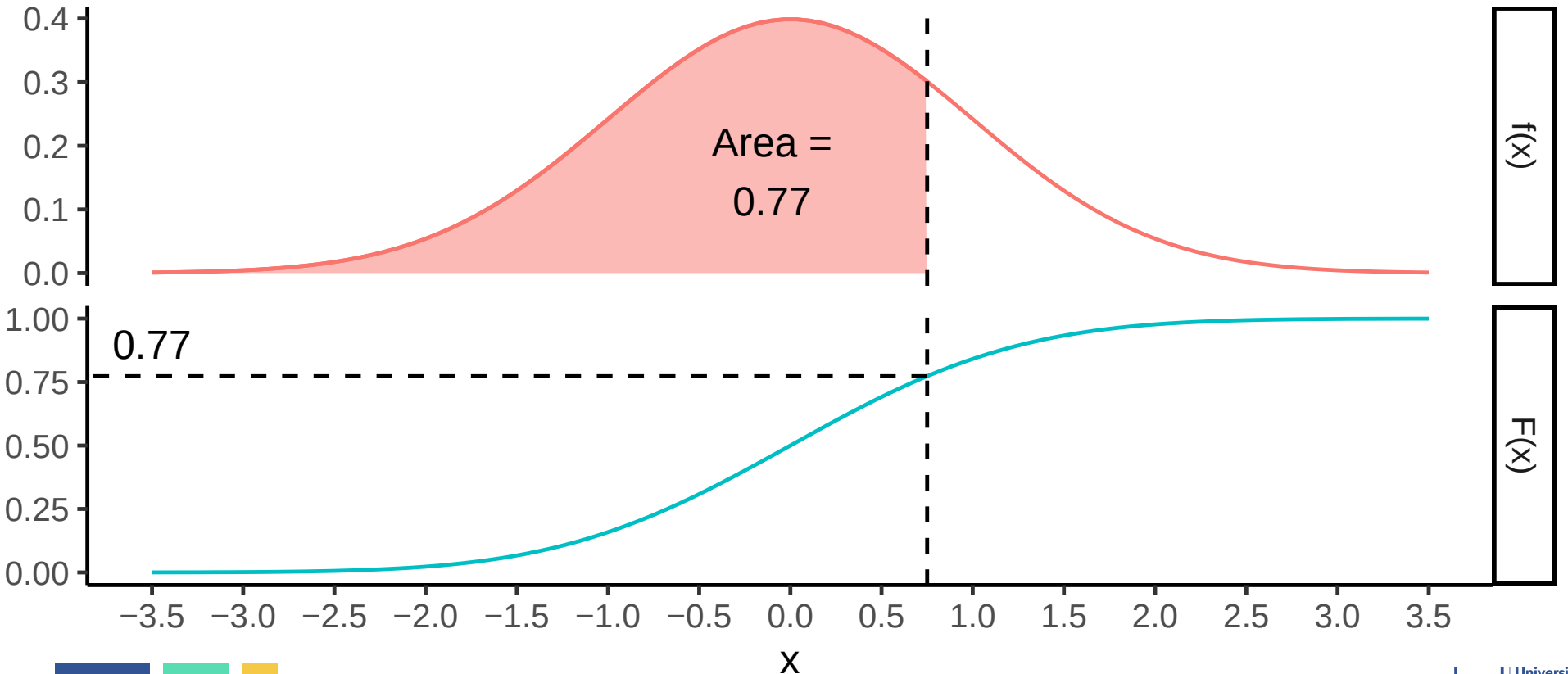
1. The bell-shaped curve is called the *density function* $[f(x)]$.
2. The curve showing areas is called the *distribution function* $[F(x)]$.

Density and distribution function

Areas under the normal curve–representing *probabilities*–are of interest.

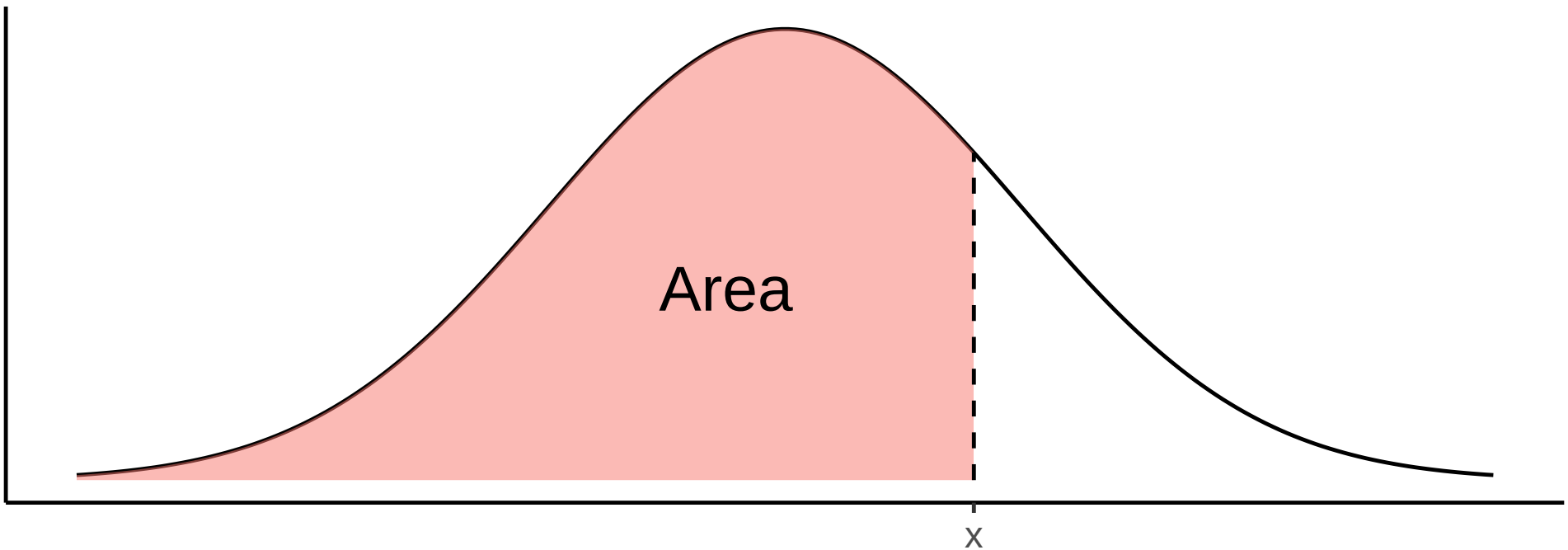
Definition 3

1. The bell-shaped curve is called the *density function* $[f(x)]$.
2. The curve showing areas is called the *distribution function* $[F(x)]$.



Probabilities

As mentioned, the value $F(x)$ represents the probability that the random variables X takes values up to and including x, i.e.



Hence, in real-world applications, we need the distribution function (rather than the density function) to calculate probabilities. (Although it's still useful to sketch the bell curve!)

Standard normal tables (cont.)

z	.00	.01	.02	.03	.04	.05	.06	.07
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340

Remarks:

- Standardise the normal variable first! $Z = (X - \mu)/\sigma$
- Only positive values of z are tabulated. The negative part may be deduced by symmetry. See exercise sheet.

Download the probability tables here: `https://haziqj.ml/stat-tables/`

Fractiles (quantiles)

As another remark, the table can be used in "reverse". To find the values of z such that $\Pr(Z < z) = p$ for some $p \in (0, 1)$, then start looking *within* the tables, and trace out the row and column digits to make up z.

z	.00	.01	.02	.03	.04	.05	.06	.07
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756

Fractiles (quantiles)

As another remark, the table can be used in "reverse". To find the values of z such that $\Pr(Z < z) = p$ for some $p \in (0, 1)$, then start looking *within* the tables, and trace out the row and column digits to make up z.

z	.00	.01	.02	.03	.04	.05	.06	.07
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756

Some important quantiles:

- $\Pr(Z < 0.00) =$
- $\Pr(Z < 0.67) = 0.75$
- $\Pr(Z < 0.84) = 0.80$
- $\Pr(Z < 1.28) = 0.90$
- $\Pr(Z < 1.64) =$
- $\Pr(Z < 1.96) =$
- $\Pr(Z < 2.57) = 0.995$
- $\Pr(Z < 3.29) = 0.9995$

Calculations

Example 4

Let us assume that the weight X of the coffee beans in one bag follows a normal distribution with mean $\mu = 500$g and SD $\sigma = 5$g.

1. What is the probability that a random coffee bag weighs at most 510g?

Calculations (cont.)

Example 4

Let us assume that the weight X of the coffee beans in one bag follows a normal distribution with mean $\mu = 500$g and SD $\sigma = 5$g.

2. What is the 95th percentile of this distribution?

Testing for normality

When dealing with data X, how do we actually know that it came from a normal distribution? We'll discuss several methods.

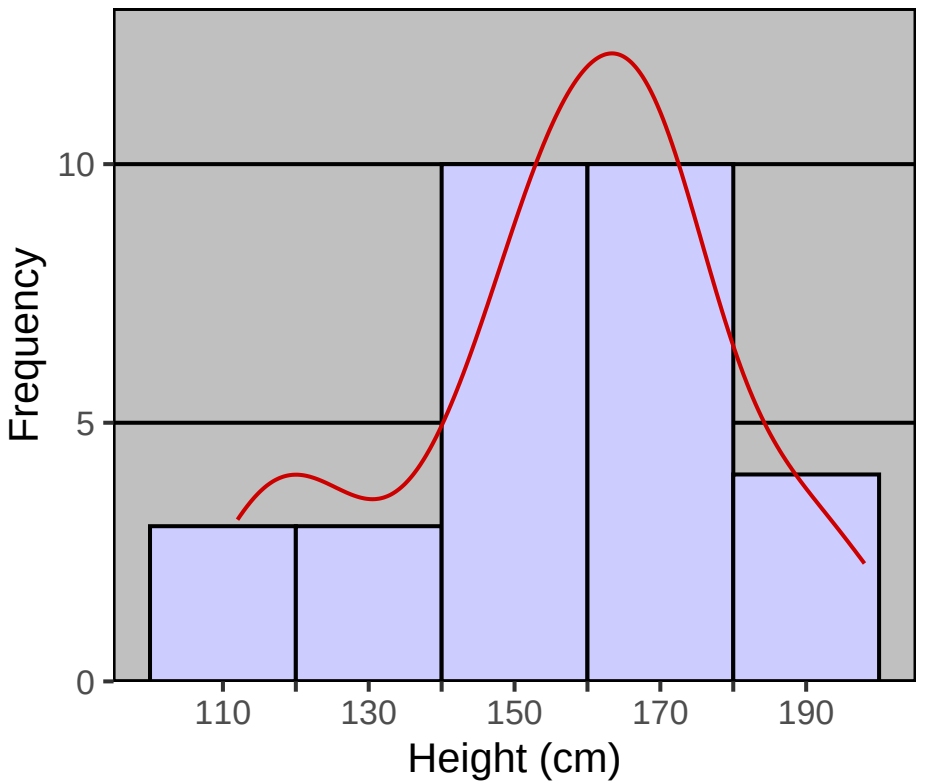
1. Check the histogram
   - Look for a symmetrical, bell-shaped appearance.
2. The average is equal to the median
   - First calculate $\bar{x}$ and M. For normal data, $\bar{x}$ and M should be very close.
3. The IQR is larger than the SD
   - Roughly speaking, for normal data, IQR $= 1.35 \times s$.
4. The 68-95-99.7 rule
   - Check that around 68% of data are within ± 1 SD of the mean, and 95% are within ± 2 SD of the mean. Majority should be within ± 3 SD.
5. QQ-plot

Example

Example 5

For the Fitness Club sample ($n = 30$), do the heights follow a normal distribution? The histogram and summary statistics are shown below.

- Mean = 157.1
- Median = 159.5
- SD = 22.1
- Q1 = 146.5
- Q3 = 170.0
- IQR = 23.5

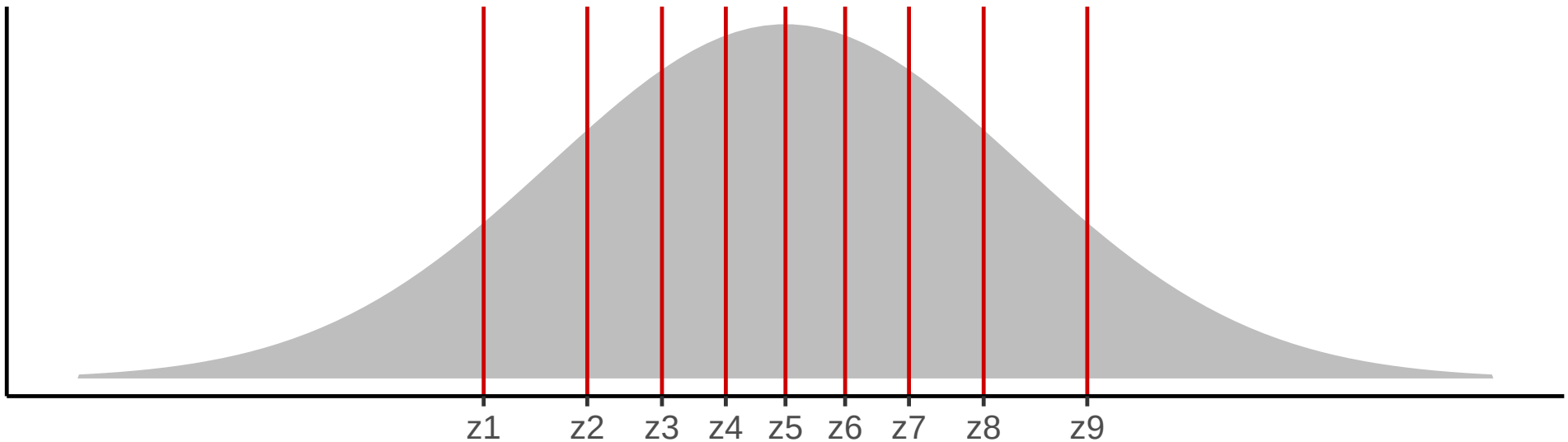


QQ-plot

Consider a subsample ($n = 9$) of the height data, arranged in order

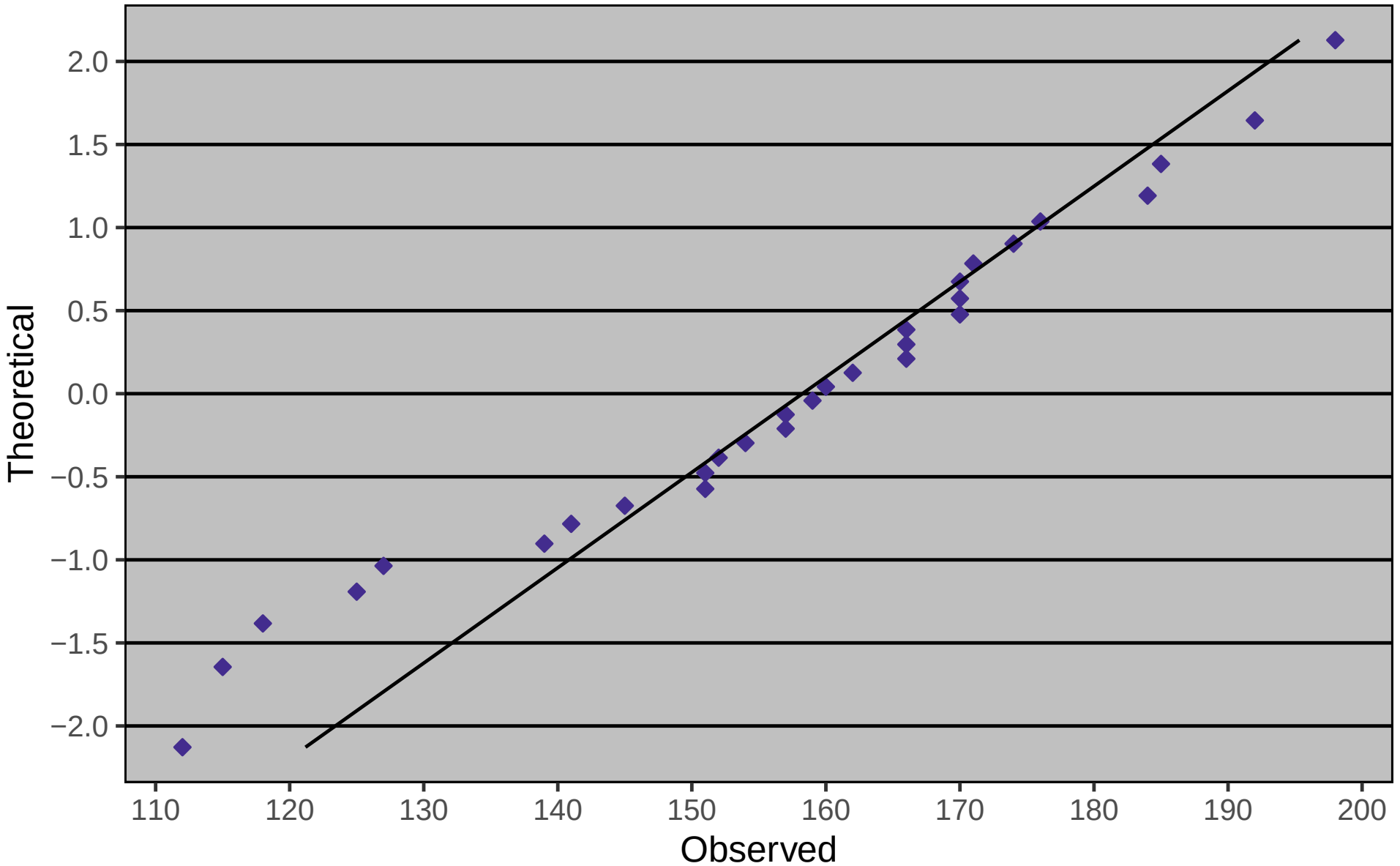
$$118\ 145\ 151\ 151\ 160\ 166\ 166\ 176\ 185$$

IDEA: If these were randomly drawn from the normal distribution, then they should each roughly belong to a section of the normal curve.



We can plot the data values against the z values, and see if they line up.

QQ-plot (cont.)



The normal distribution

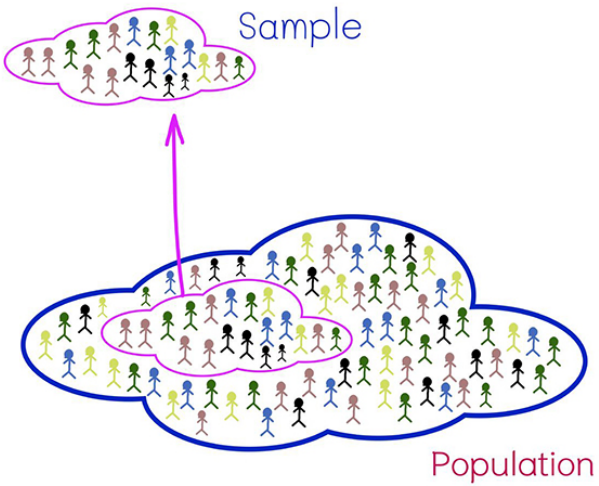
Statistical inference

Statistical inference

Back to the coffee conundrum:

> *If the weight of coffee beans in the bags is randomly normal, what can we conclude about the (average) weight?*

- We can check for normality using the techniques discussed prior.
- However, often μ and σ are **unknown** in the population.



Definition 6

The estimators for the unknown population μ and σ are the sample mean and SD respectively, i.e.

i. $\hat{\mu} = \bar{x}$
ii. $\hat{\sigma} = s$

This estimate is random

Suppose that I purchase 25 bags of coffee, and use the sample mean to estimate μ. Here are some numbers.

499 501 503 492 511 503 487 501 503 501 498 500 497 493
498 511 492 499 501 500 500 496 505 494 506

$\bar{x} = 499.67$

This estimate is random

Suppose that I purchase 25 bags of coffee, and use the sample mean to estimate μ. Here are some numbers.

499 501 503 492 511 503 487 501 503 501 498 500 497 493
498 511 492 499 501 500 500 496 505 494 506

$\bar{x} = 499.67$

I could repeat this another time...

497 499 504 508 499 504 502 505 502 496 496 507 503 497
495 495 497 505 504 499 502 499 506 503 504

$\bar{x} = 501.04$

This estimate is random

Suppose that I purchase 25 bags of coffee, and use the sample mean to estimate μ. Here are some numbers.

499 501 503 492 511 503 487 501 503 501 498 500 497 493
498 511 492 499 501 500 500 496 505 494 506

$\bar{x} = 499.67$

I could repeat this another time...

497 499 504 508 499 504 502 505 502 496 496 507 503 497
495 495 497 505 504 499 502 499 506 503 504

$\bar{x} = 501.04$

And another...

501 500 493 504 496 501 495 504 501 497 503 498 505 500
493 500 499 502 487 495 501 506 492 492 501

$\bar{x} = 498.65$

This estimate is random

Suppose that I purchase 25 bags of coffee, and use the sample mean to estimate μ. Here are some numbers.

499 501 503 492 511 503 487 501 503 501 498 500 497 493
498 511 492 499 501 500 500 496 505 494 506

$\bar{x} = 499.67$

I could repeat this another time...

497 499 504 508 499 504 502 505 502 496 496 507 503 497
495 495 497 505 504 499 502 499 506 503 504

$\bar{x} = 501.04$

And another...

501 500 493 504 496 501 495 504 501 497 503 498 505 500
493 500 499 502 487 495 501 506 492 492 501

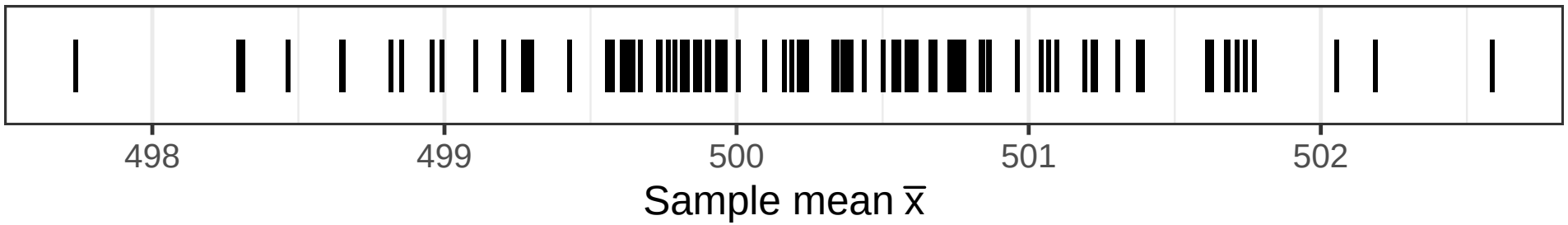
$\bar{x} = 498.65$

And yet another...

499 502 497 496 504 506 503 496 503 499 500 496 496 507
504 499 501 499 499 502 497 500 503 489 503

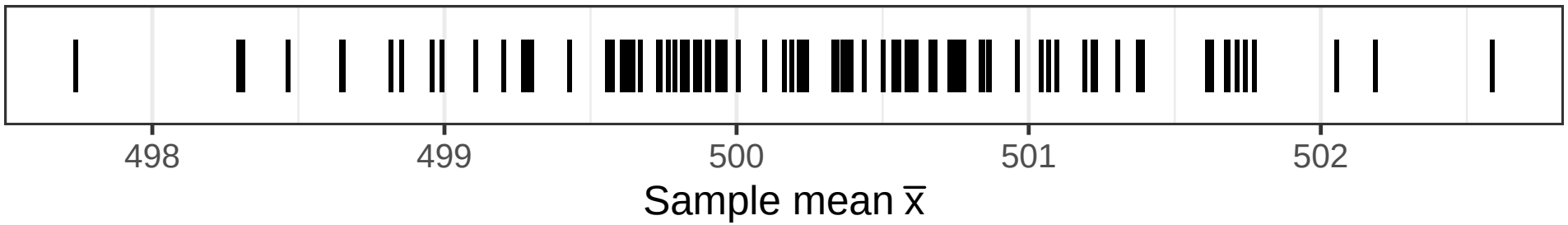
$\bar{x} = 499.96$

Repeating this 100 times



- Everytime we take a *random* sample, we get a *random* $\bar{x}$.
- Perhaps instead of providing a *point* estimate, we can give an *interval* estimate instead.
- What are the set of values which have a high chance, 95% say, of containing the true value μ?

Repeating this 100 times



- Everytime we take a *random* sample, we get a *random* $\bar{x}$.
- Perhaps instead of providing a *point* estimate, we can give an *interval* estimate instead.
- What are the set of values which have a high chance, 95% say, of containing the true value μ?

Definition 7 (Standard error)

The variability of the sample mean is captured by the quantity known as the *standard error*, defined

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

Confidence interval

Definition 8 (Confidence interval for the mean)

The 95% confidence interval for the mean is obtained using the formula

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

Some remarks

- In practice, σ is unknown, so is replaced by the sample version s.
- The value '1.96' corresponds to the upper and lower 2.5% points of the normal distribution–suggesting that the distribution of $\bar{x}$ is also normal!
- When n is small (say $n < 20$), typically the t-distribution is used instead. More on this in Chapter 6.

Example

Example 9

Calculate a 95% CI for the mean height of all kids in the population (Fitness Club).

145 151 118 166 160 151 166 185 176 125 152 127 112 157 115 174 171 141 166 162 157 139 159 170 198 192 154 170 184 170

- $\bar{x} = 157.10$ cm
- $s = 22.06$ cm
- $n = 30$