Normal Curve

Modelling data | Normal Model

STAT5002

The University of Sydney

Feb 2025

THE UNIVERSITY OF SYDNEY

Data Modelling

Topic 3: Normal Curve

What is the Normal Curve? And what does it have to do with sample mean?

Topic 4: Linear Model

How can we describe the relationship between two variables? When is a linear model appropriate?

Outline

Data story

The normal curve

Area under normal curves

Special properties

Quantiles

Data story

How likely is it to find an elite netball goal player in Australia?

ABC NEWS

Just In Australia World Business Sport Science Arts Analysis Fact Check

Print Email Facebook Twitter More

Tall netballers put through their paces at the Australian Institute of Sport

By Jonathon Gul

Updated 14 Jun 2015, 11:49am

Tall young netballers from around the country have been gathering at the Australian Institute of Sport (AIS) in Canberra to develop their agility and speed on court.

A total of 10 goal shooters, goal attacks, goal keepers and goal defenders, who were all over 189cm in height, and all younger than 25 years.

Former Australian netball team member and AIS Centre of Excellence coach Jenny Borlase is running the camp.

She said the group represented the future of Australian netball, and was hopeful the young players would go on to compete in the national competition.

"Tall goal shooters and tall goal keepers are much more a part of the scene than when I was playing 15 years ago," she said.

"We recognise that in Australia the game of netball is changing, we want to remain competitive and maintain a competitive advantage."

PHOTO: Height can be a strong advantage when shooting a goal in a netball game. (ABC News: Ian Cutmore)

ABC News

Statistical Thinking

"A total of 10 goal players (shooters, keepers, attacks and defenders) … were **all over 189cm** in height".

How could you investigate the proportion of Australian women who are over 189cm in height (potential elite goal players)?

- Collect the heights of Female students in the unit. For example, we have the data collected from "Statistical Thinking with Data" (MATH1005) in 2022 S2

Then we have two options:

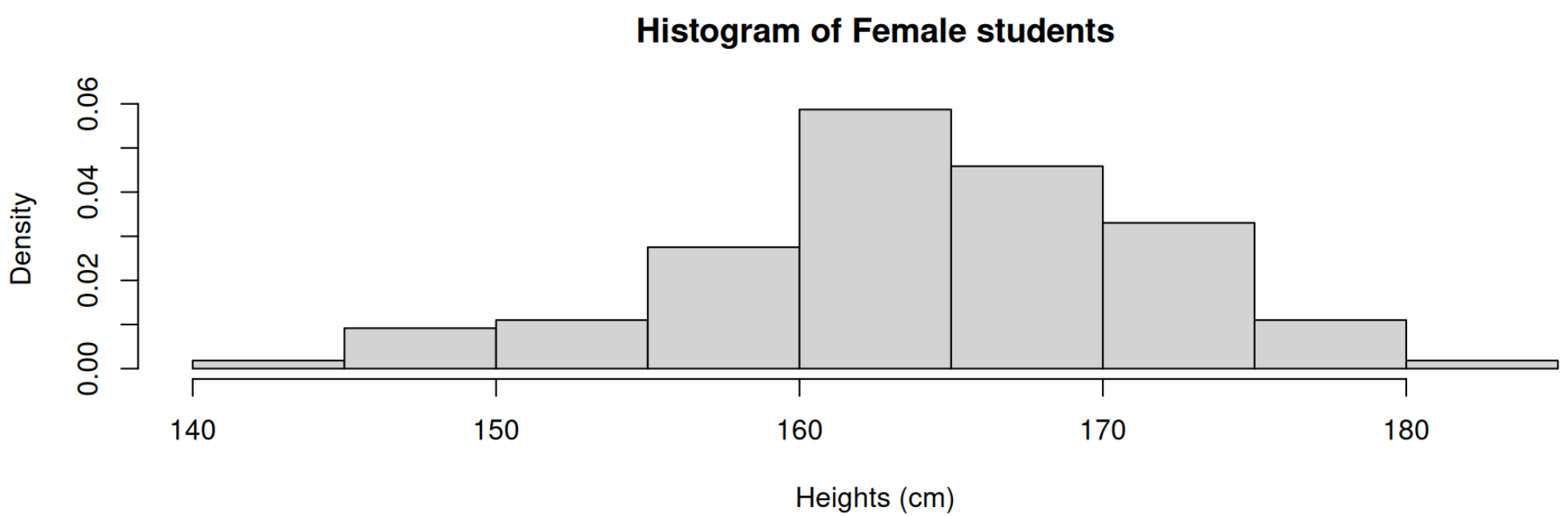
- Use the data to represent the population
- Use the data to create a model for the population

Investigation: Data from MATH1005, 2022 S2

```r
math1005 = read.csv("data/math1005_cleaned.csv", header = T)
FemaleHeights = math1005$Height[math1005$Gender == "Female"]
FemaleHeights = na.omit(FemaleHeights)
length(FemaleHeights)  # There were 109 female students
```

```
[1] 109
```

```r
hist(FemaleHeights, main = "Histogram of Female students", xlab = "Heights (cm)",
    freq = F)
```



Numerical Investigation

```
1 mean(FemaleHeights)
```

```
[1] 164.8633
```

```
1 sd(FemaleHeights)
```

```
[1] 7.516324
```

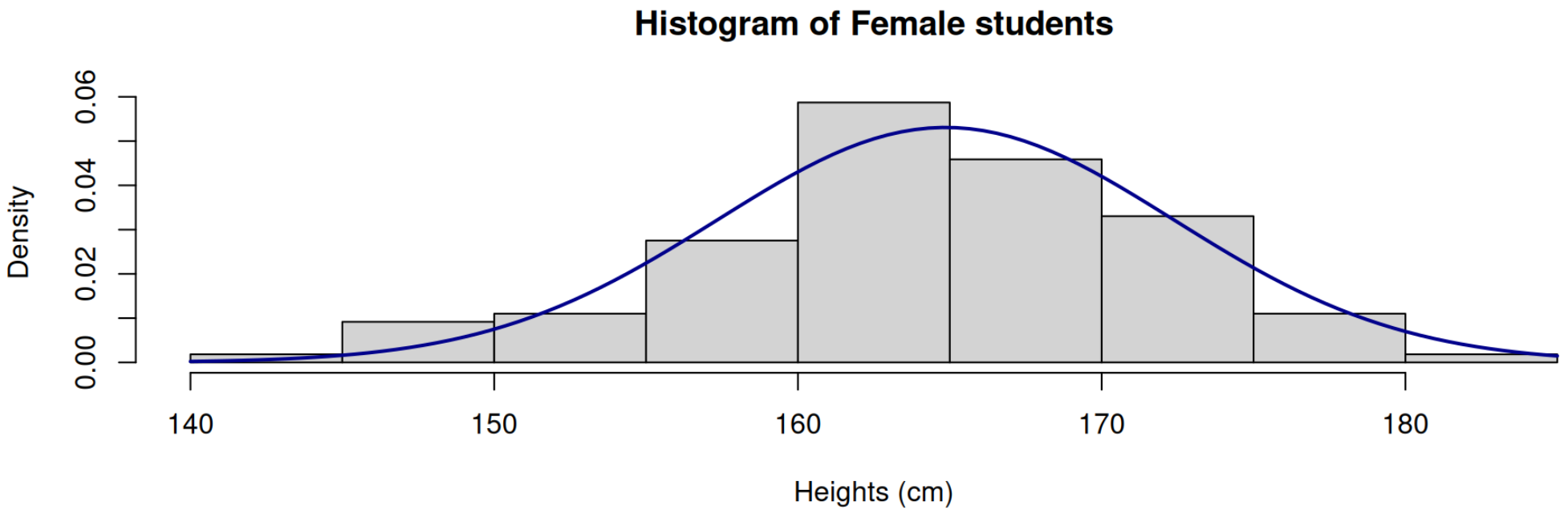
```
1 # sum(...) counts the number of FemaleHeights > 189
2 sum(FemaleHeights > 189)/length(FemaleHeights)
```

```
[1] 0
```

> **Note**
>
> How many students could be elite goal players?

- In this sample, none! But we know there are women in Australia taller than 189cm...

Approximate density-scale histogram



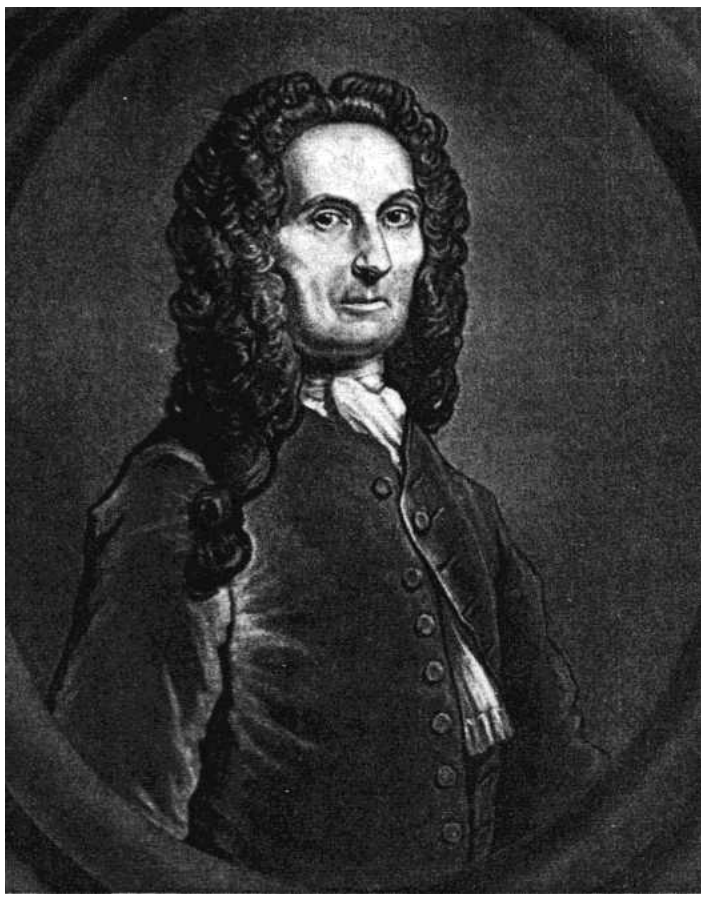
We can draw a smooth curve to approximate the **density-scale** histogram. This curve may extend beyond the range of observed data in the sample to allow us to answer the research question.

How would you describe its shape?

- Fairly symmetric and bell-shaped. Is there something special about this curve?

Normal curve

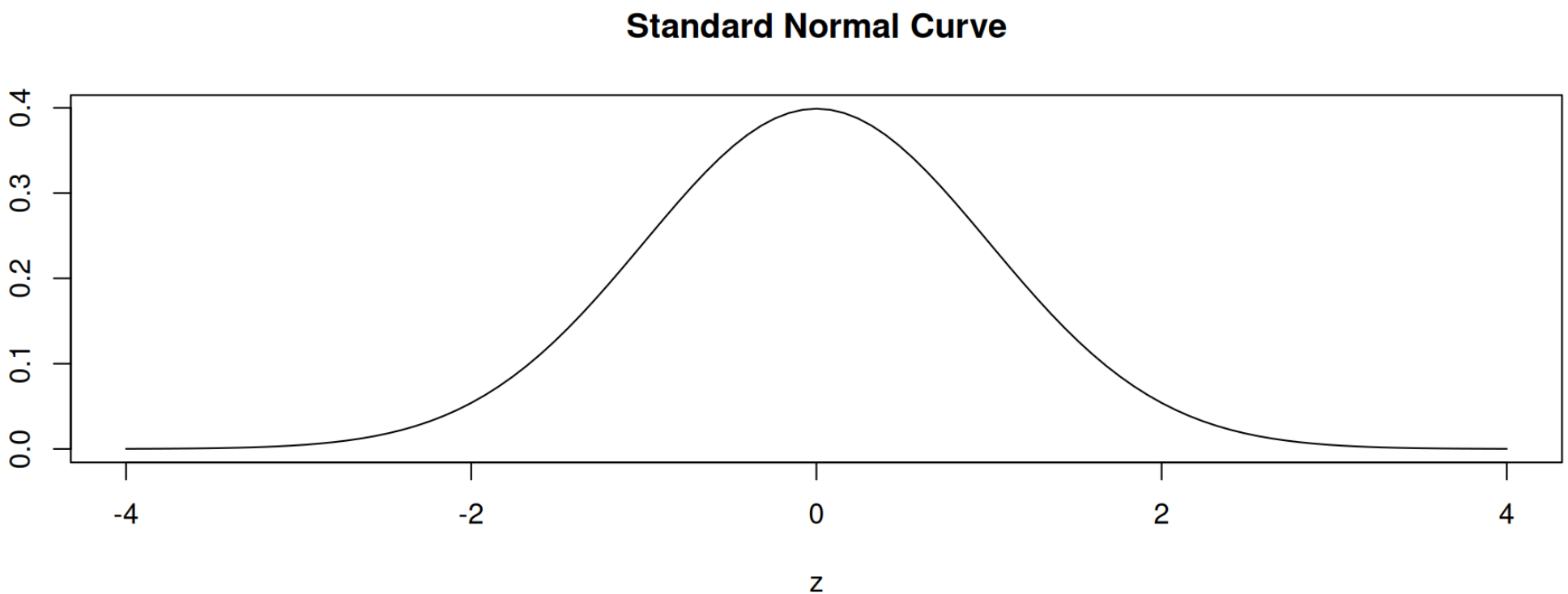
The Normal curve was defined around 1720 by Abraham de Moivre, also famous for the beautiful de Moivre's formula.



- The Normal curve approximates many **natural phenomena**.
- The Normal curve can model data caused by combining a **large number of independent observations**. (Coming up in a future lecture after introducing probability)
- Many of its properties can be obtained using elementary single variable calculus.

General & Standard Normal curves

- The **General** Normal Curve (X) has any mean and SD. Caution: It is denoted by N(mean, **Variance**), where **Variance** = SD^2.
- The **Standard** Normal Curve (Z) has mean 0 and SD 1. Short: $N(0,1)$



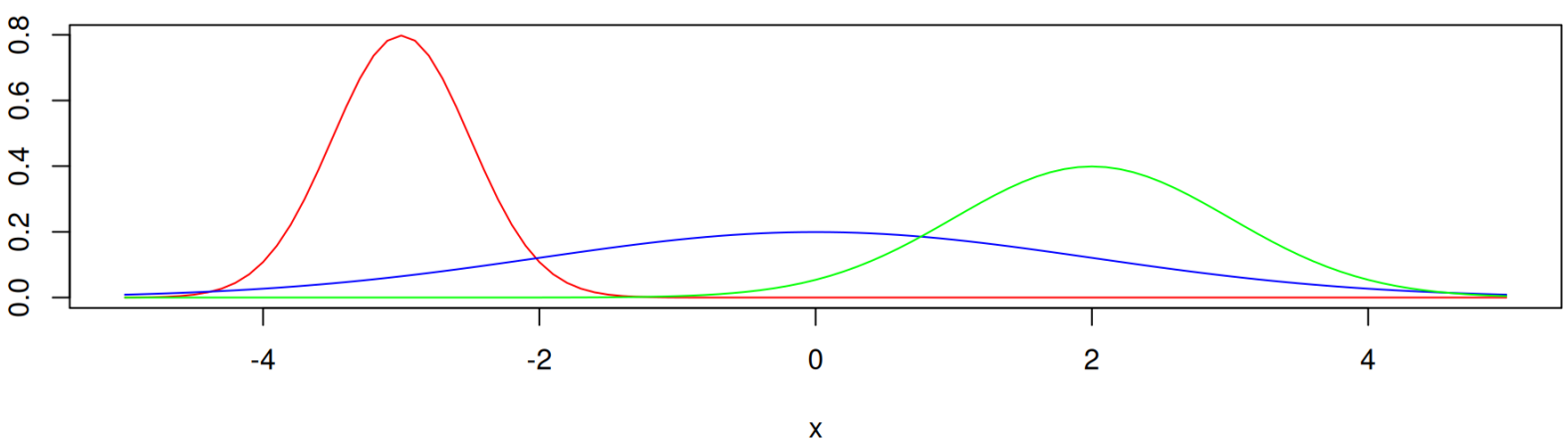
The Normal curve formula

The **general normal curve** can be described by the formula

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } x \in (-\infty, \infty)$$

where we can control the shape by (μ, σ):

- μ is the mean, or the centre of the curve
- σ is the standard deviation, or the spread of the curve.



Area under normal curves

Area under a normal curve

The area under any general normal curve $N(\mu, \sigma^2)$, bounded by some interval (a, b), is given by

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

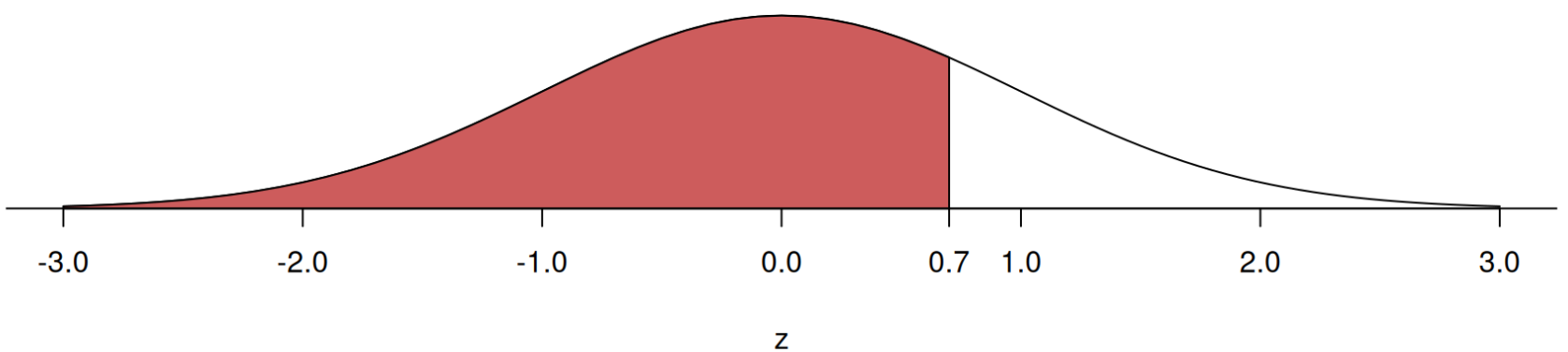
- The total area under the normal curve (between $a = -\infty$ and $b = \infty$) is 1.
- X denotes data following a general normal curve with mean μ and SD σ.
- $P(a < X < b)$ denotes the proportion of data falling into the interval (a, b).
- We will later use this notation also for probability and random variables.

Simplification: Area under the standard normal curve

We start with some data Z modeled by the standard normal curve $N(0,1)$. As $\mu = 0$ and $\sigma = 1$, the proportion of data falling into the interval (a, b) is

$$P(a < Z < b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz.$$

For example, the proportion of data that is 0.7 or lower is given by the area up to 0.7.



But how to calculate this?

Method 1: Integration

By its definition, we could use integration:

$$P(Z < 0.7) = \int_{-\infty}^{0.7} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

But this does not have a closed form.

Method 2: Normal tables (not for assessment)

This is the old way. We table the values of the integral.

TABLE 1. **Lower tail areas of the Standard Normal distribution (CDF)** The point tabulated is $\Phi(z) = P(Z \leq z)$, where $Z \sim N(0, 1)$.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177

Method 3: Use R

- The `pnorm(x)` command works out the **lower tail** area, it gives

$$P(Z < x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

- The `pnorm(x,lower.tail=F)` command works out the **upper tail** area, it gives

$$P(Z > x) = \int_{x}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$$

- We also have

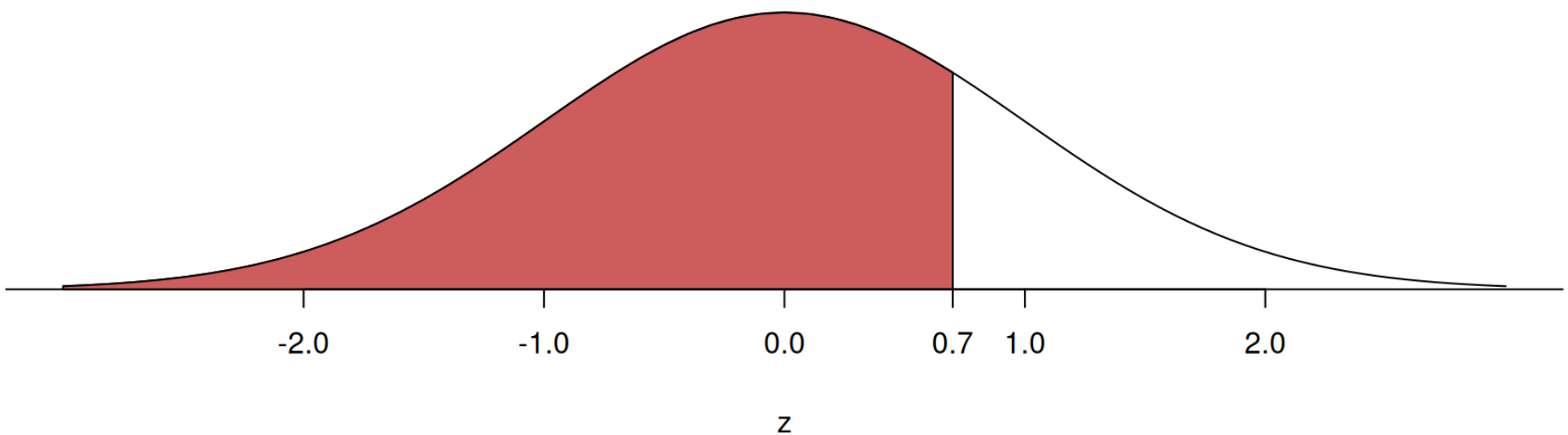
$$P(Z > x) = 1 - P(Z < x) \quad \text{or} \quad \text{upper tail area} = 1 - \text{lower tail area}$$

- It is useful to sketch the normal curve and the relevant area ... and then use R.

Lower tail

What proportion of data is 0.7 or lower?

$$P(Z < 0.7) \approx 0.76$$



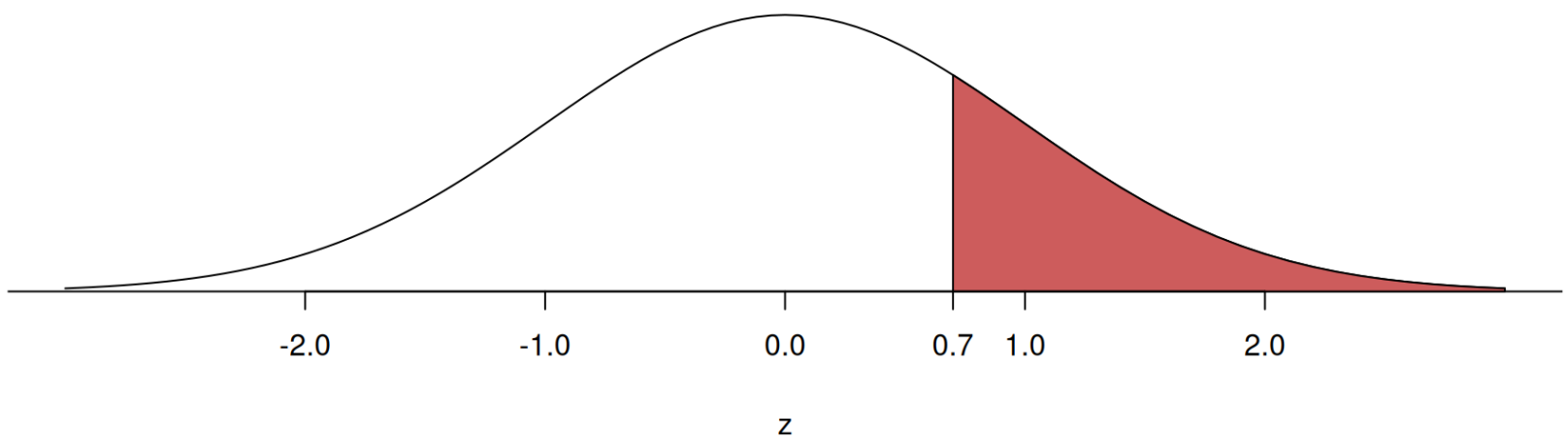

```
1 pnorm(0.7)
```

```
[1] 0.7580363
```

Upper tail

What proportion of data is 0.7 or higher?

$$P(Z > 0.7) \approx 0.24$$




```
1 pnorm(0.7, lower.tail = F)
```

```
[1] 0.2419637
```

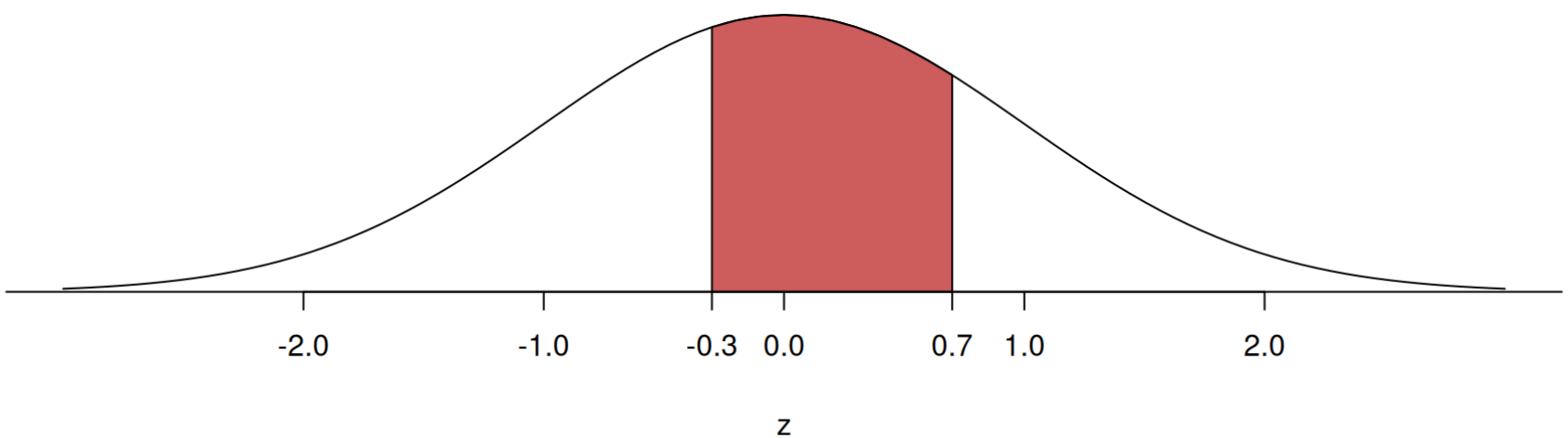
```
1 1 - pnorm(0.7)  # alternative way
```

```
[1] 0.2419637
```

Interval

What proportion of data is between -0.3 and 0.7?

$$P(-0.3 < Z < 0.7) = \underbrace{P(Z < 0.7)}_{\int_{-\infty}^{0.7} f(z)dz} - \underbrace{P(Z < -0.3)}_{\int_{-\infty}^{-0.3} f(z)dz} \approx 0.38$$




```
1 pnorm(0.7) - pnorm(-0.3)
```

```
[1] 0.3759478
```

Area under general normal curves

Heights of female students in MATH1005

```
1 mean(FemaleHeights)
```

```
[1] 164.8633
```

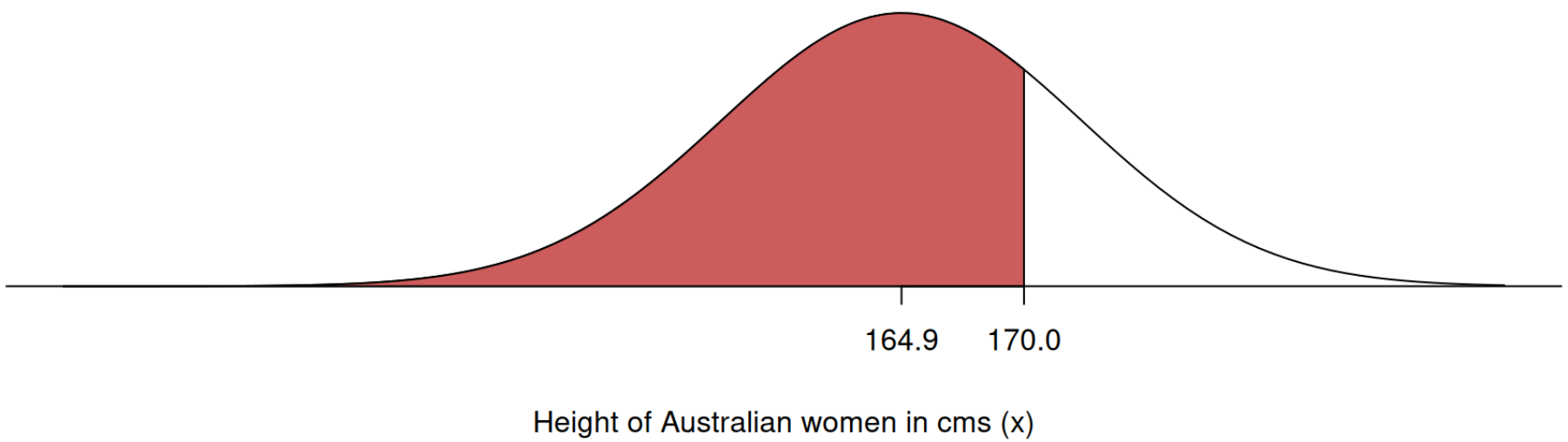
```
1 sd(FemaleHeights)
```

```
[1] 7.516324
```

- The heights of female students in MATH1005 has a mean of 164.9cm and a standard deviation of 7.52cm.
- Now we can model the heights of all Australian women with a normal curve with mean 164.9cm and standard deviation of 7.5cm.

Lower tail

Suppose the heights of Australian women follow a normal distribution with mean 164.9cm and sd 7.5cm. What proportion of women will have height less than 170cm?

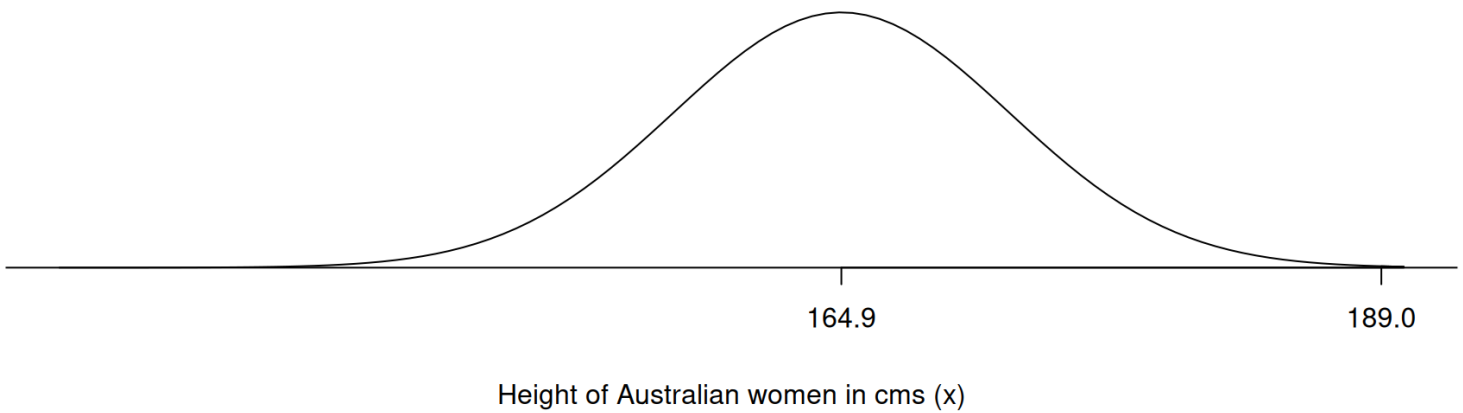



```r
m = mean(FemaleHeights)
s = sd(FemaleHeights)
pnorm(170, m, s)  #pnorm(x,mean,sd)
```

```
[1] 0.7528247
```

Upper tail

What proportion of women will have height greater than 189cm? How likely is to find an elite netball goal player in Australia?




```r
m = mean(FemaleHeights)
s = sd(FemaleHeights)
pnorm(189, m, s, lower.tail = FALSE)  #upper tail, pnorm(x,mean,sd)
```

```
[1] 0.0006608243
```

```r
1 - pnorm(189, m, s)  # 1 - lower tail
```

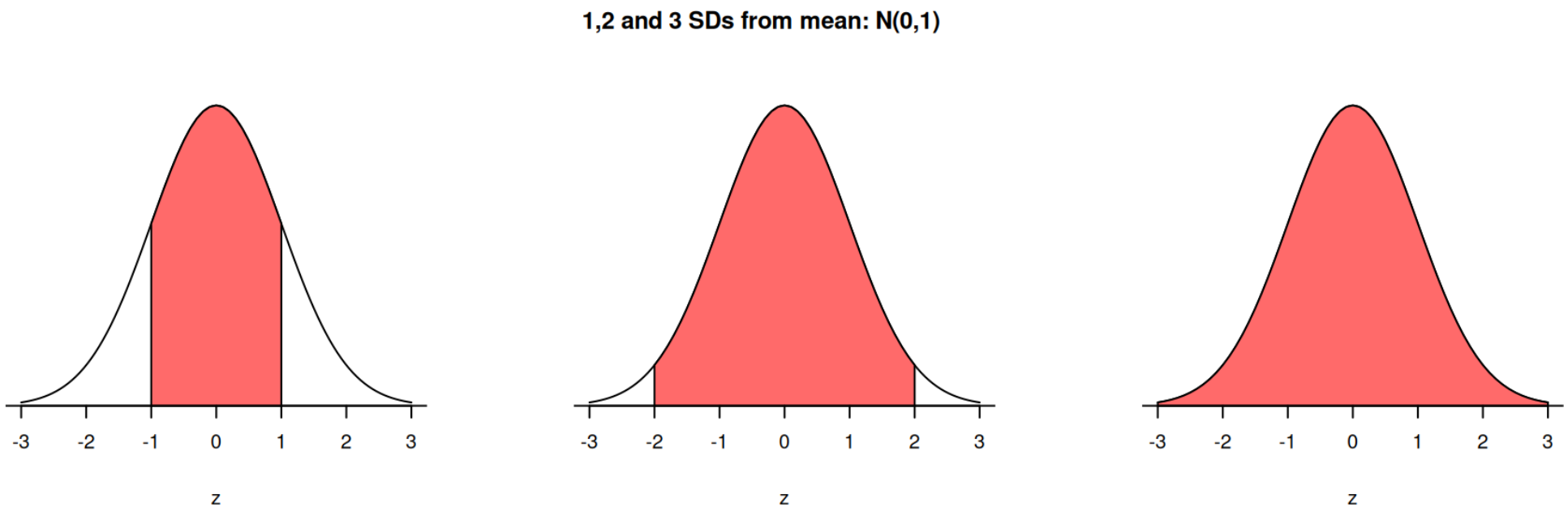
```
[1] 0.0006608243
```

Properties of the normal curve

68% 95% 99.7% Rule

All normal curves satisfy the “68%-95%-99.7% rule”:

- The area **1** SD out from the mean in both directions is **0.68** (68%).
- The area **2** SDs out from the mean in both directions is **0.95** (95%).
- The area **3** SDs out from the mean in both directions is **0.997** (99.7%).



Under a normal curve, it has a low chance (0.3%) to have data points that fall more than 3 SD away from the mean.

Rescaling

Any general normal curve can be rescaled into the standard normal curve.

Consider data X following a general normal curve $N(\mu, \sigma^2)$. For any point on this normal curve, recall that the standard unit (or z score) is how many standard deviations that point is above (+) or below (-) the mean.

$$\text{standard unit} = \frac{\text{data point - sample mean}}{\text{sample SD}} \quad \text{or} \quad z = \frac{x - \mu}{\sigma}$$

- The standard units give the relative location of a data point on the standard normal curve.
- The proportion under a general normal curve $P(X < a)$ is equivalent to the proportion under the standard normal curve $P(Z < \frac{a-\mu}{\sigma})$

The proportion of data modelled by $N(\mu, \sigma^2)$ falling below a is

$$P(X < a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Apply the change of variable (standardisation) $z = \frac{x-\mu}{\sigma}$

$$\int_{-\infty}^{a} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \int_{-\infty}^{\frac{a-\mu}{\sigma}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}z^2} \frac{dx}{dz} dz$$

where

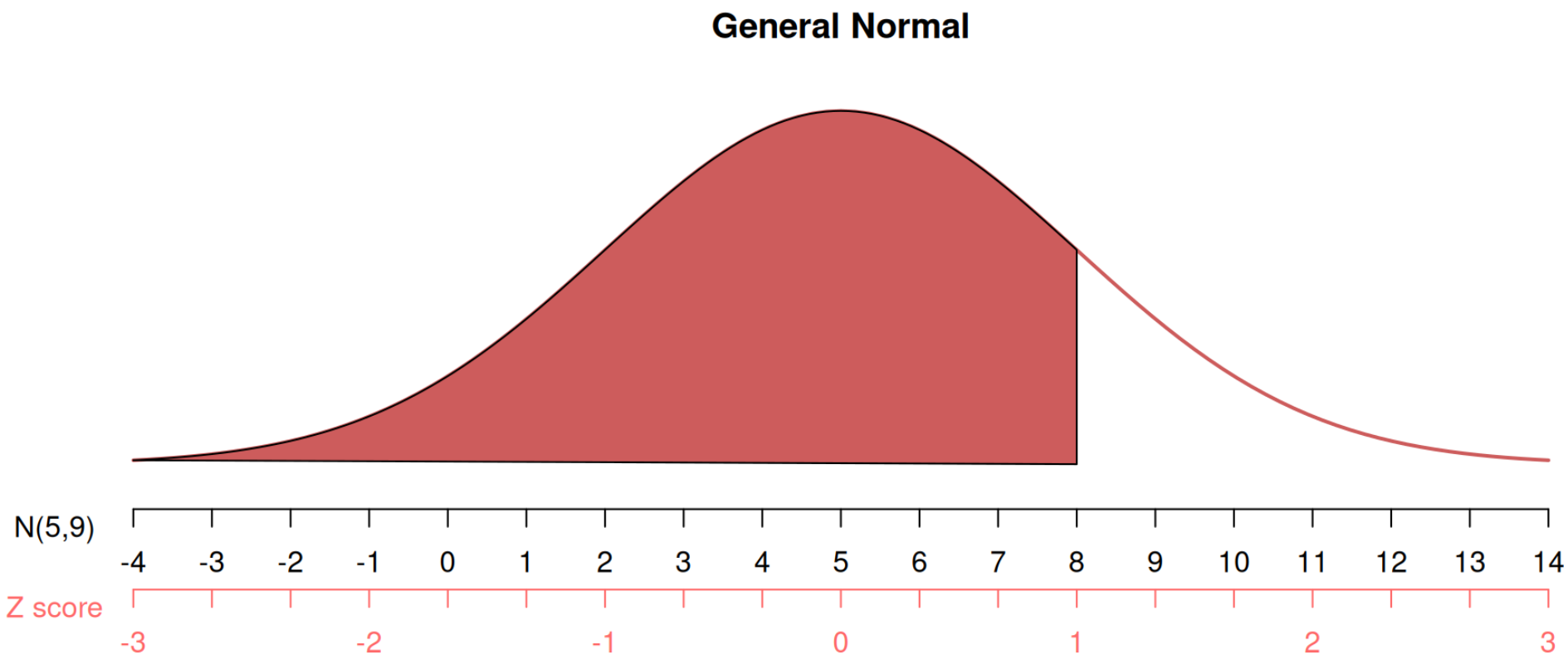
$$z = \frac{x-\mu}{\sigma} \quad \Longrightarrow \quad x = \sigma z + \mu \quad \Longrightarrow \quad \frac{dx}{dz} = \sigma$$

so the proportion simplifies to

$$P(X < a) = \int_{-\infty}^{\frac{a-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = P\left(Z < \frac{a-\mu}{\sigma}\right)$$

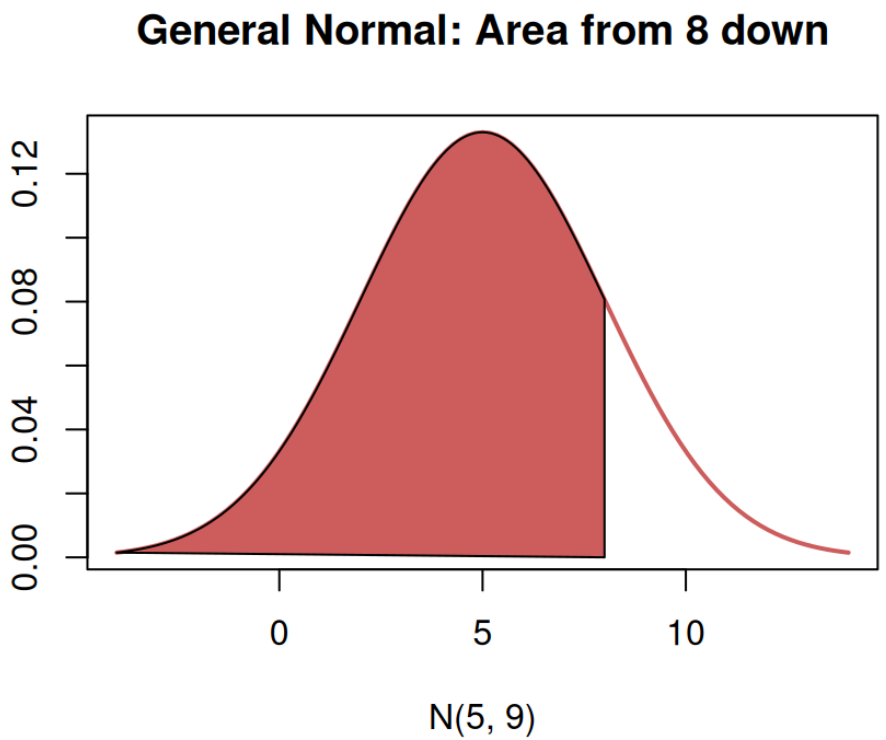
which is the proportion of data modelled by $N(0, 1)$ falling below $\frac{a-\mu}{\sigma}$.

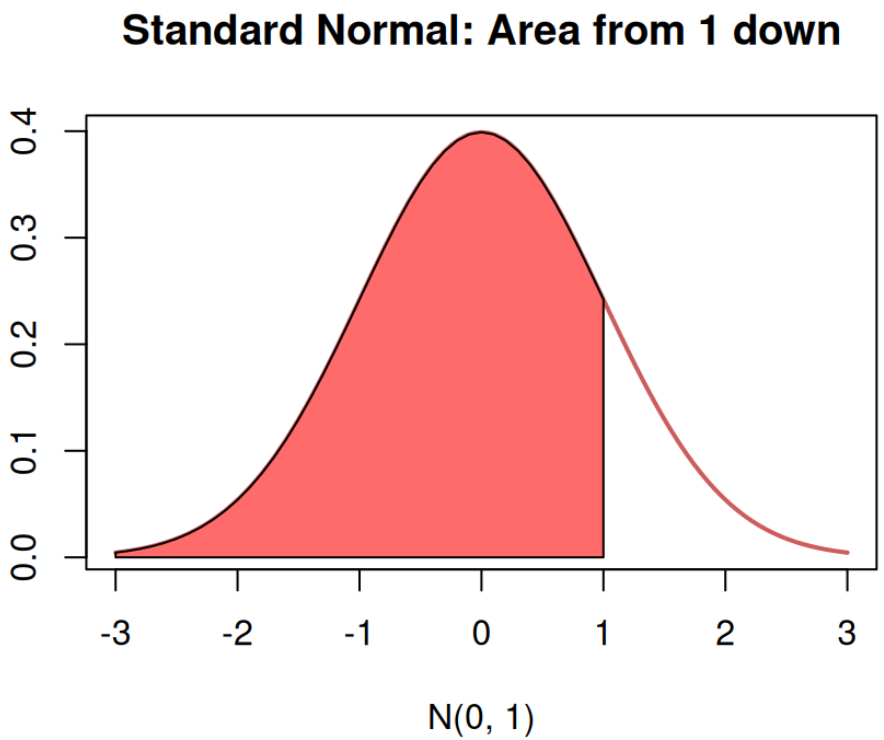
Example 1



- Consider the point = 8.
- So the z score of the point is $\frac{8-5}{3} = 1$.

The following 2 areas are of the same size.





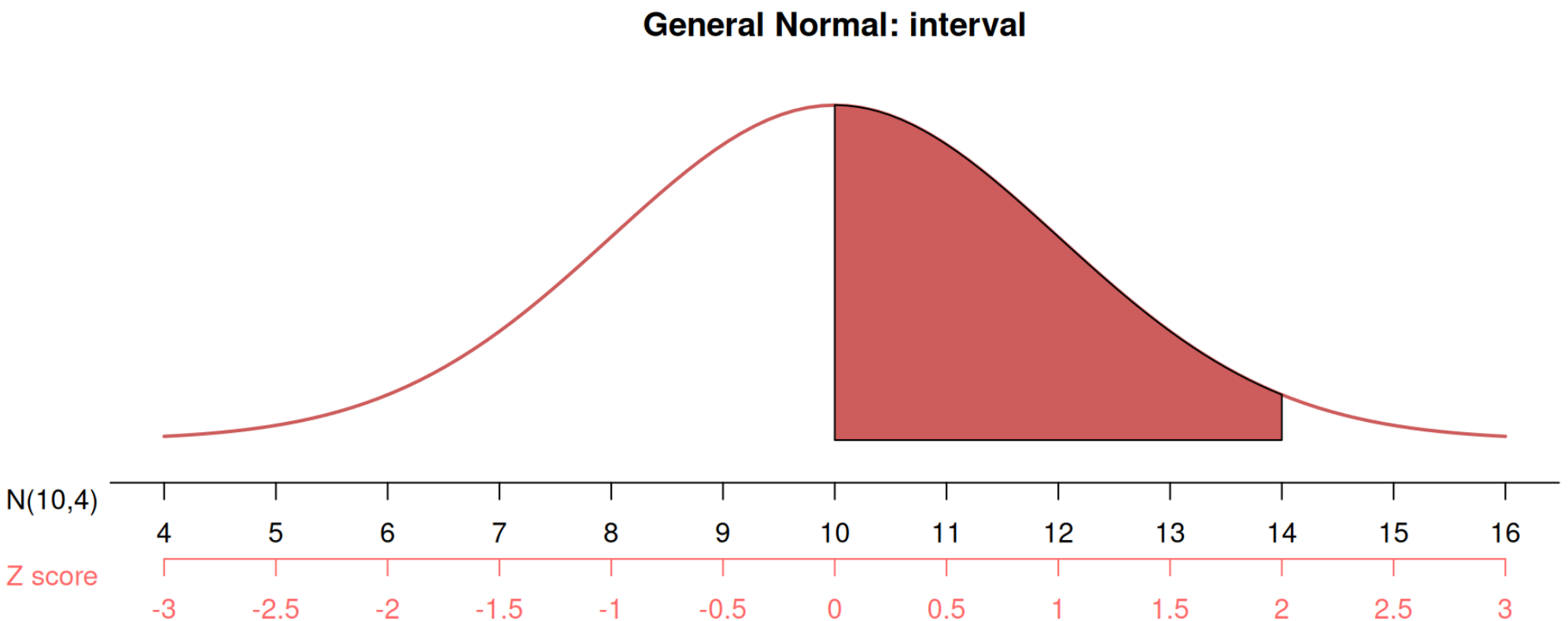

```
1 pnorm(8, 5, 3)
```

```
[1] 0.8413447
```

```
1 pnorm(1)
```

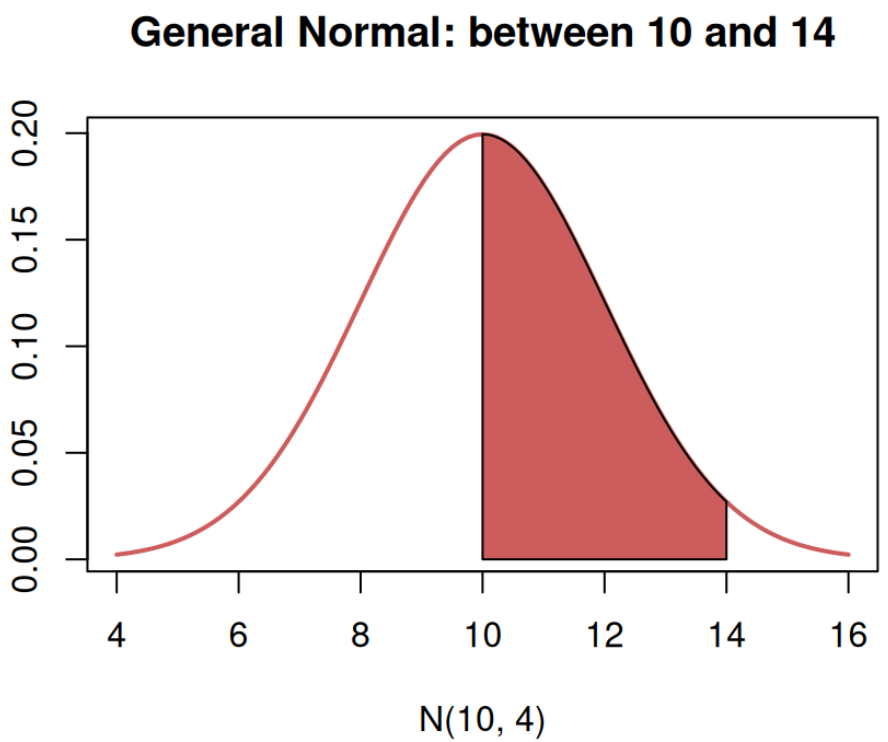
```
[1] 0.8413447
```

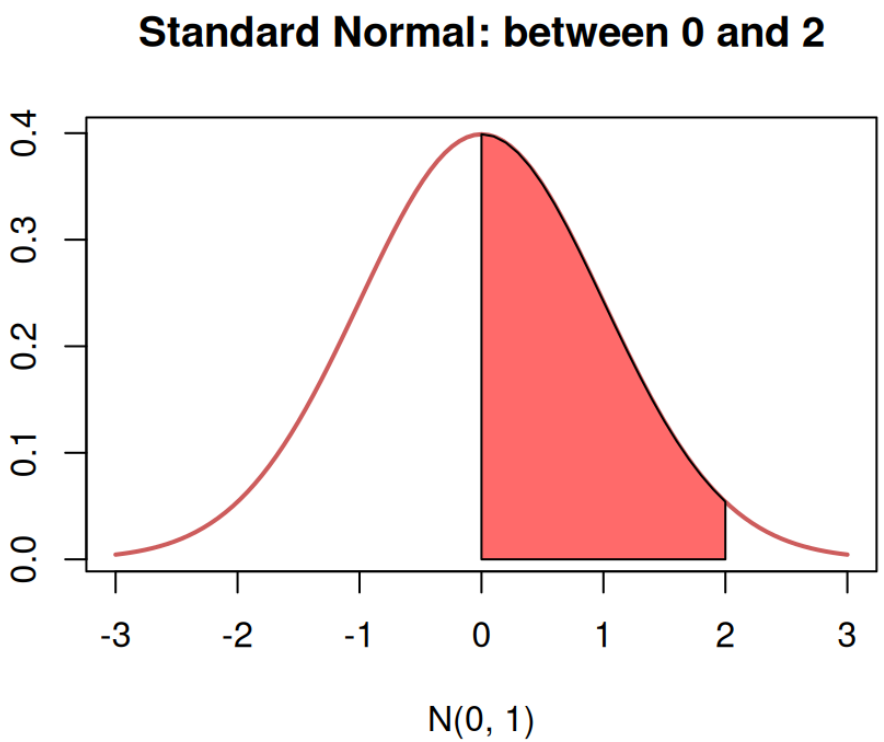
Example 2



- Here the lower point is 10 and the upper point is 14.
- So the z scores are $z_1 = \frac{10-10}{2} = 0$ and $z_2 = \frac{14-10}{2} = 2$.

The following 2 areas are of the same size.






```
1 pnorm(14, 10, 2) - pnorm(10, 10, 2)
```

```
[1] 0.4772499
```

```
1 pnorm(2) - pnorm(0)
```

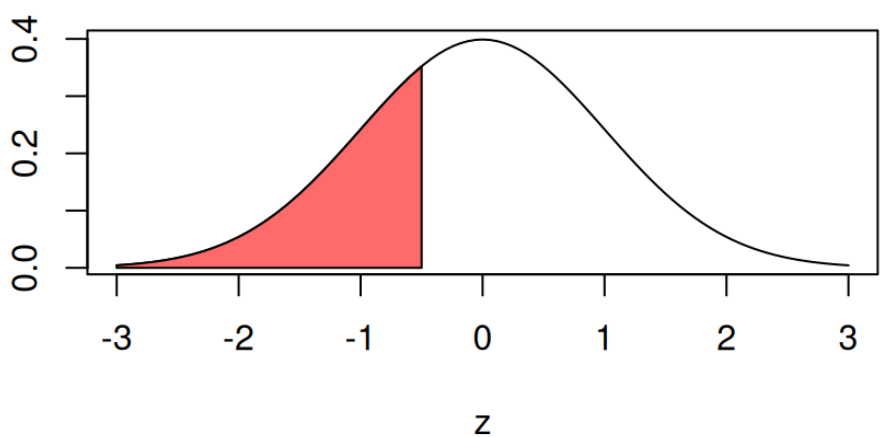
```
[1] 0.4772499
```

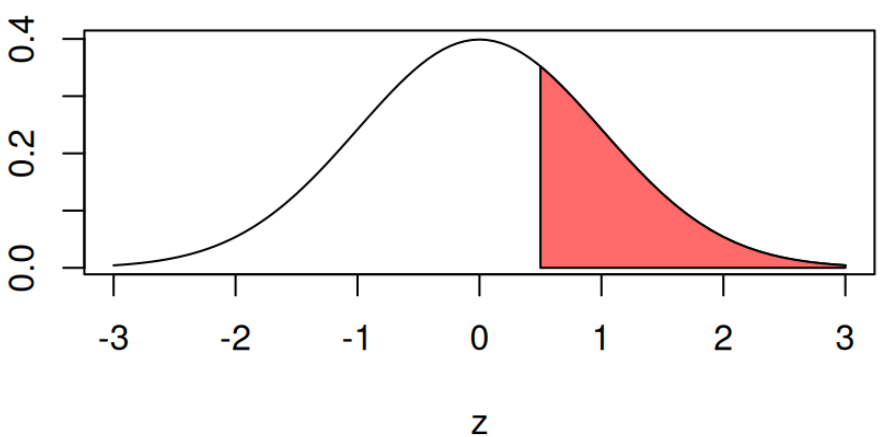
The normal curve is symmetric about the mean

If Z follows the standard normal curve $N(0,1)$, then

$$P(Z < -a) = P(Z > a)$$

The red areas below are of the same size (where $a = 0.5$).





More generally, if X follows a general normal curve $N(\mu, \sigma^2)$, then

$$P(X < \mu - a) = P(X > \mu + a)$$

```r
mu = 10
sigma = 2
a = 2
pnorm(mu - a, mu, sigma)  # lower tail
```

```
[1] 0.1586553
```

```r
pnorm(mu + a, mu, sigma, lower.tail = F)  # upper tail
```

```
[1] 0.1586553
```

Calculate the quantiles of normal curves using R

The function `pnorm()` finds "the proportion of data X following a normal curve falling below the value a", we are also interested in

- What is the quantile Q such that p% of the data X falling below the value Q?

Similar to the proportion, there is no close-form solution for the quantiles of normal curves. We can calculate the quantiles using `qnorm(x, mu, sigma)` in R.

```r
mu = 10
sigma = 2
qnorm(0.7, mu, sigma)  # 70-percentile of N(10, 4)
```

```
[1] 11.0488
```

```r
qnorm(0.5, mu, sigma)  # 50-percentile (or the median) of N(10, 4)
```

```
[1] 10
```

```r
qnorm(0.7)  # 70-percentile of the standard normal N(0, 1)
```

```
[1] 0.5244005
```

```r
qnorm(0.5)  # 50-percentile (or the median) of N(0, 1)
```

```
[1] 0
```

Summary

- The Normal curve naturally describes many histograms, and so can be used in modelling data.
- It can be described by the mean and the variance (SD^2).
- Area under normal curves and `pnorm`.
- It has many useful properties, including the 68/95/99.7% rule.
- Any general normal can be rescaled into a standard normal.
- The normal curve is symmetric about the mean.
- Quantiles and `qnorm`.