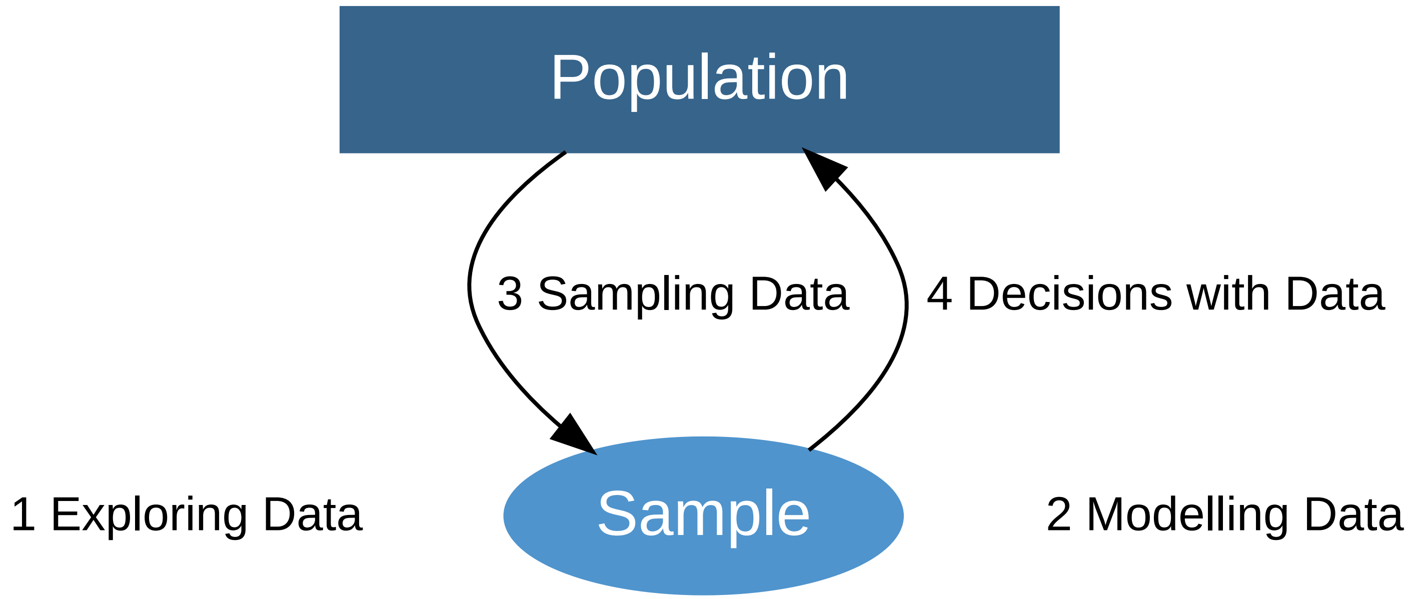
Normal Curve

Modelling data | Normal Model

Unit Overview



Module2 Modelling Data

Normal Model

What is the Normal Curve? How can we use it to model data?

Linear Model

How can we describe the relationship between 2 variables? When is a linear model appropriate?

Normal Curve

Data Story

How likely is 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

“A total of 10 goal shooters, goal keepers, goal attacks and goal defenders … were **all over 189cm** in height”.

Statistical Thinking

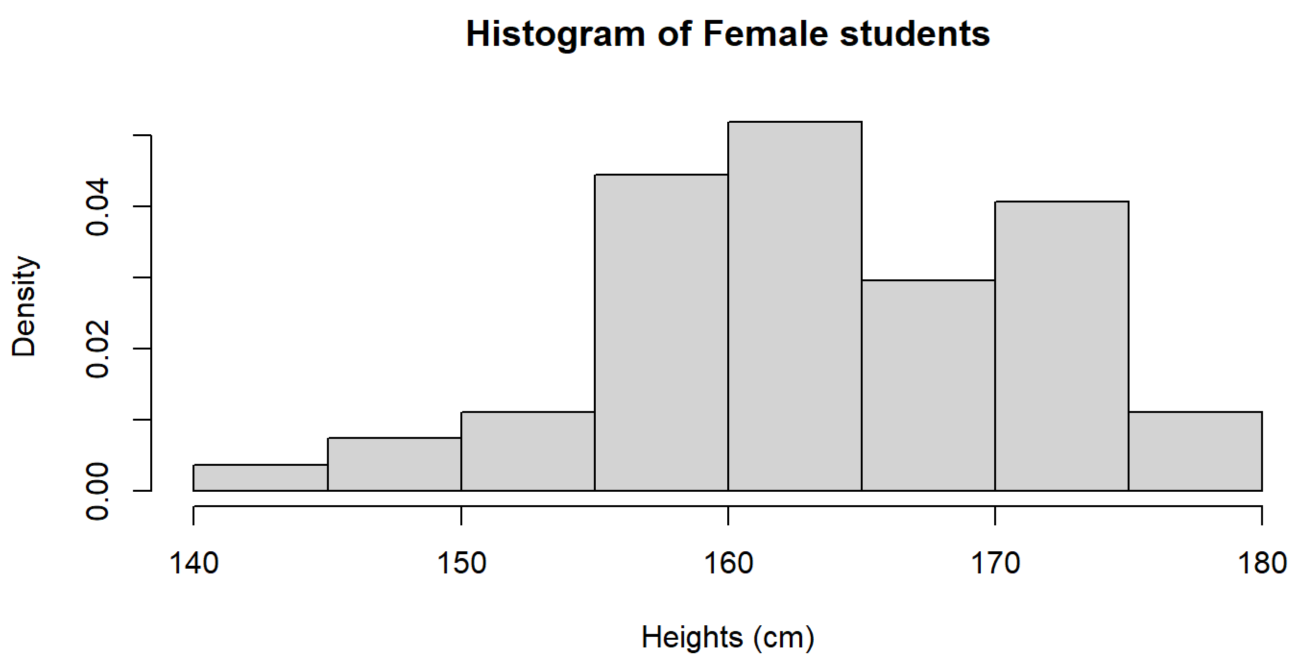
How could you investigate the proportion of Australian women who are over 189cm in height?

- Collect the heights of Australian female students in DATA1001 and produce a histogram: `studentheights.csv`.
- Investigate the ABS data which tells us that the mean is 161.8 (cm): `ABSFemaleHeights.xlsx`.

Investigation1: Data from DATA1001

The following data is the heights of 53 female students from DATA1001 in 2018.

```
data = read.csv("data/studentheights.csv", header = T)
hist(data$heights, main = "Histogram of Female students", xlab = "Heights (cm)",
    freq = F)
```



```
mean(data$heights)
```

```
## [1] 163.6204
```

```
sd(data$heights)
```

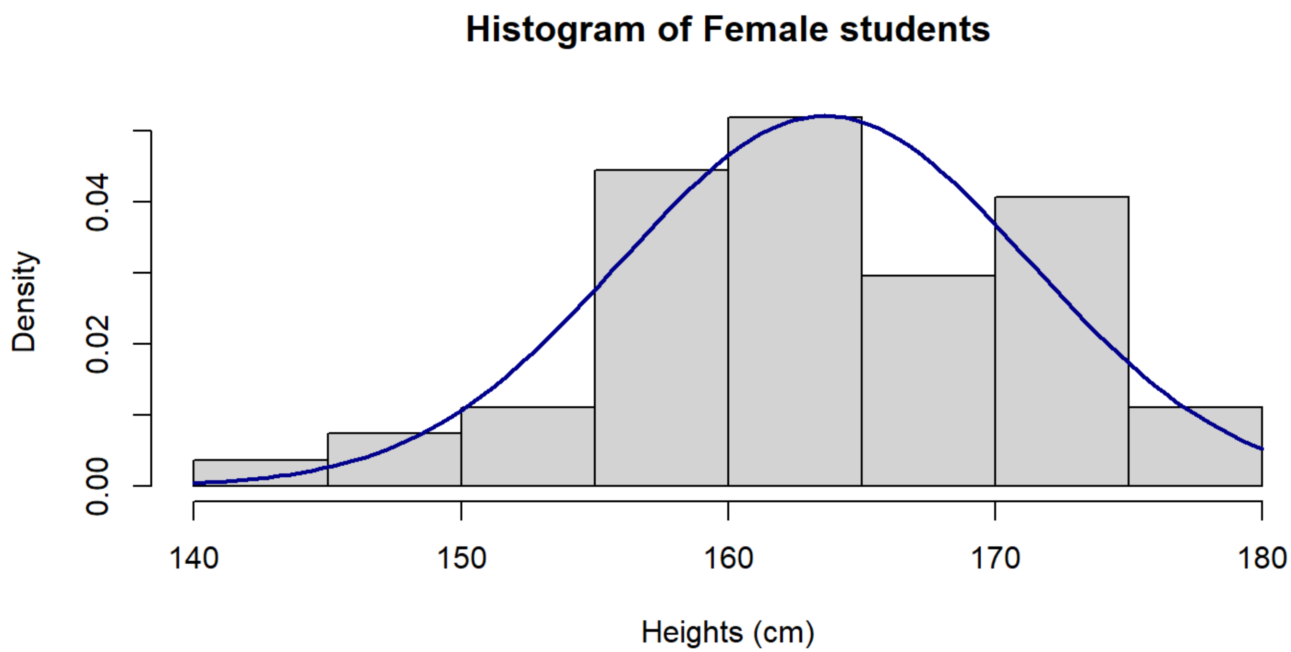
```
## [1] 7.657661
```

```
length(data$heights[data$heights > 189])/length(data$heights)
```

```
## [1] 0
```

How many students could be goal players?

- In this sample, none!



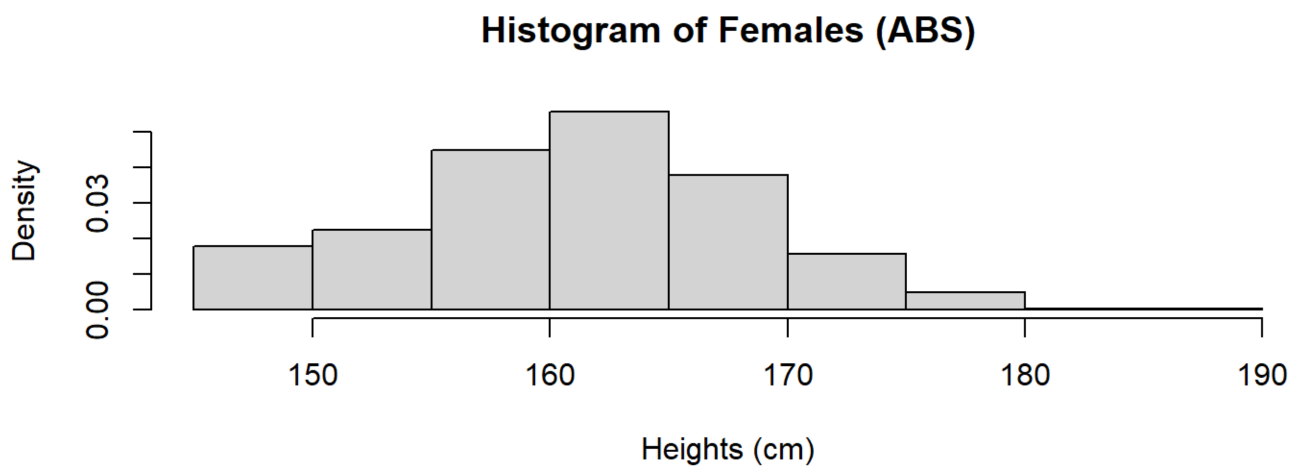
If we drew a smooth curve as an approximation of the histogram, how would you describe it’s shape?

- Fairly symmetric and bell-shaped.

Investigation2: Data from ABS

From the Australian Health Survey (2011), we get the following (simulated) data for the heights of 7,154 women.

```
data1 = read.csv("data/ABSFemaleHeights.csv", header = T, sep = "")  # Unfortuntately this is summarised data, not raw.
set.seed(1)
data2 = c(sample(145:155, 7154 * 0.166, replace = T), sample(155:160, 7154 * 0.215,
    replace = T), sample(160:165, 7154 * 0.286, replace = T), sample(165:170, 7154 *
    0.209, replace = T), sample(170:175, 7154 * 0.09, replace = T), sample(175:180,
    7154 * 0.029, replace = T), sample(180:190, 7154 * 0.005, replace = T))  # Simulation of data1 (Ext)
hist(data2, main = "Histogram of Females (ABS)", xlab = "Heights (cm)", freq = F)
```



```
mean(data2)
```

```
## [1] 161.8665
```

```
sd(data2)
```

```
## [1] 7.610036
```

```
data2[data2 > 189]
```

```
## [1] 190 190 190
```

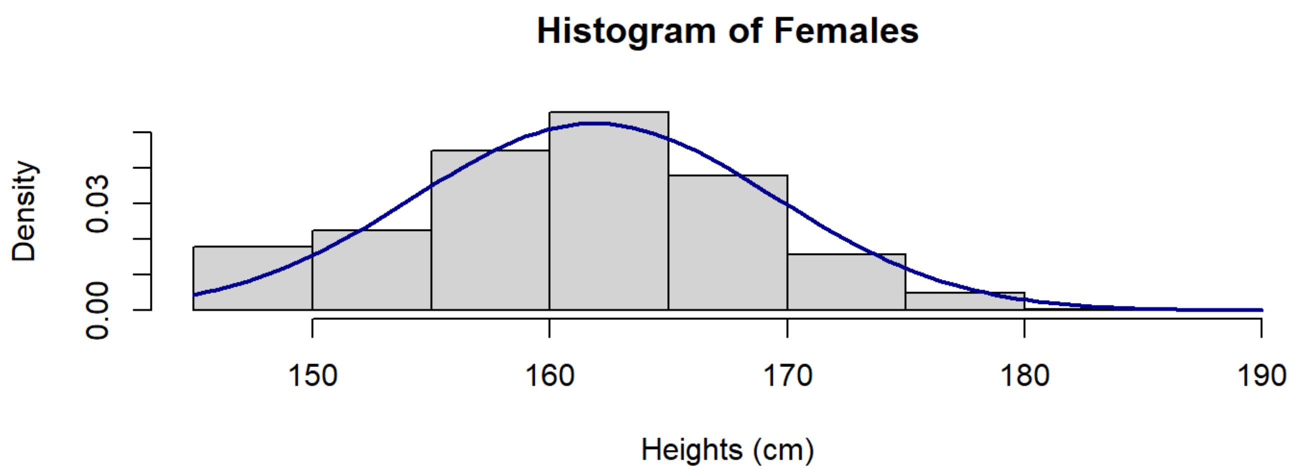
```
length(data2[data2 > 189])/length(data2)
```

```
## [1] 0.0004195217
```

How many women could be goal players?

- In this sample, there are 3 women, which is 0.04%.

```r
hist(data2, main = "Histogram of Females", xlab = "Heights (cm)", freq = F)
m2 = mean(data2)
sd2 = sd(data2)
curve(dnorm(x, mean = m2, sd = sd2), col = "darkblue", lwd = 2, add = TRUE)
```



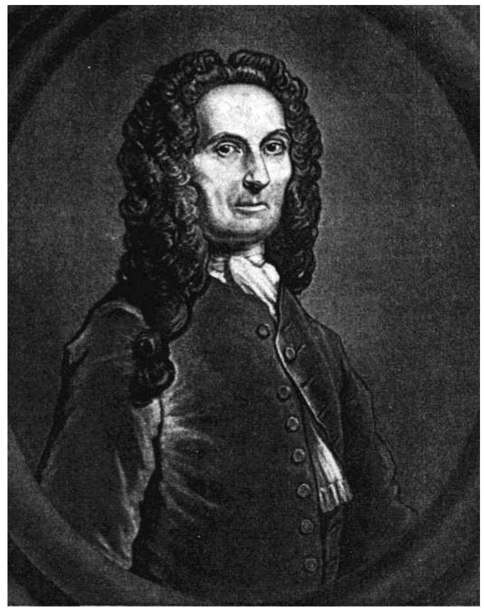
If we drew a smooth curve as an approximation of the histogram, how would you describe it's shape?

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

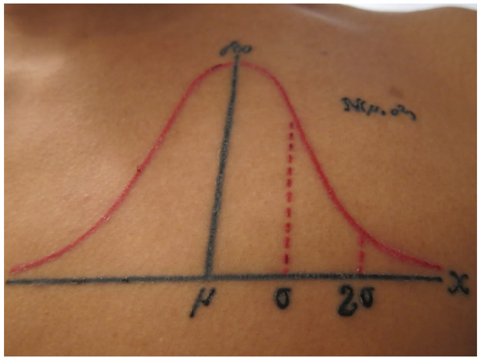
The Normal Curve

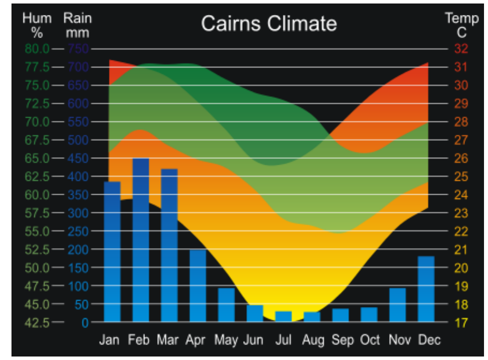
Origins of the Normal Curve

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



Why is the Normal curve famous?

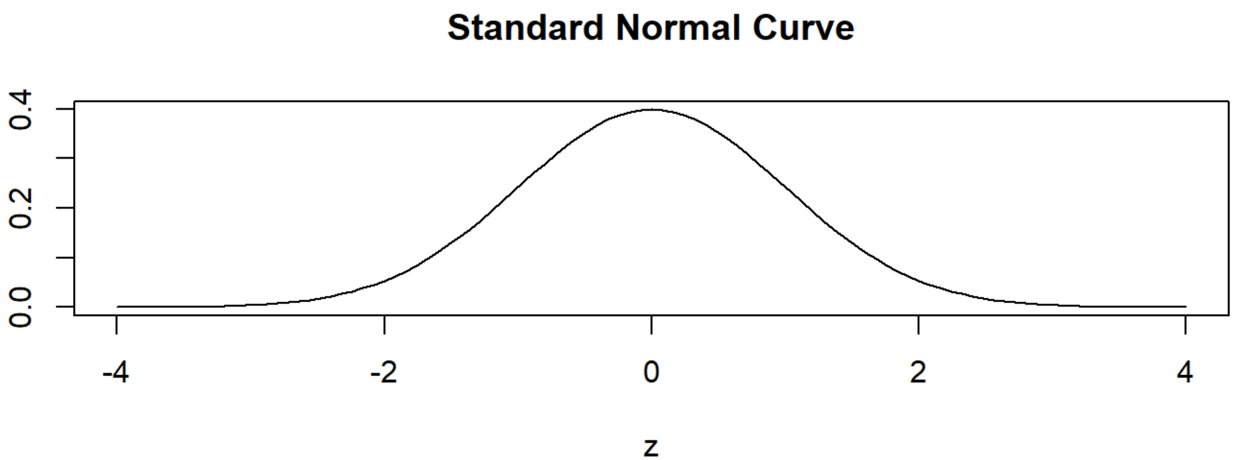


- The Normal curve approximates many **natural phenomemon**.
- The Normal curve can model data caused by combining a **large number of independent variables**. (Coming up in Normal Approximation.)

General & Standard Normal curves

- The **General** Normal Curve (X) has any mean and SD.
- The **Standard** Normal Curve (Z) has mean 0 and SD 1.



The Normal curve formula (Ext)

- It turns out the Normal curve has a simple formula, although you won't need to use it directly.

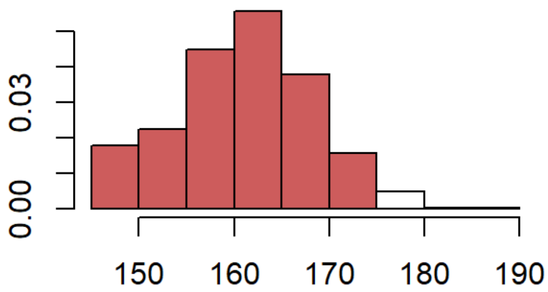
- The formula for the General Normal Curve is

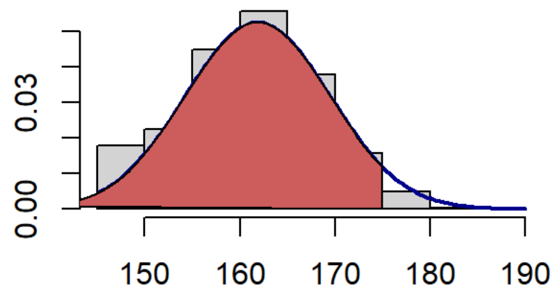
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } x \in (-\infty, \infty)$$

where μ and σ are the (population) mean and SD respectively.

Approximating histogram by a Normal curve

If the Normal curve seems to fit the histogram, then we can use the **area under the Normal curve** as an approximation to the **area under the histogram**.



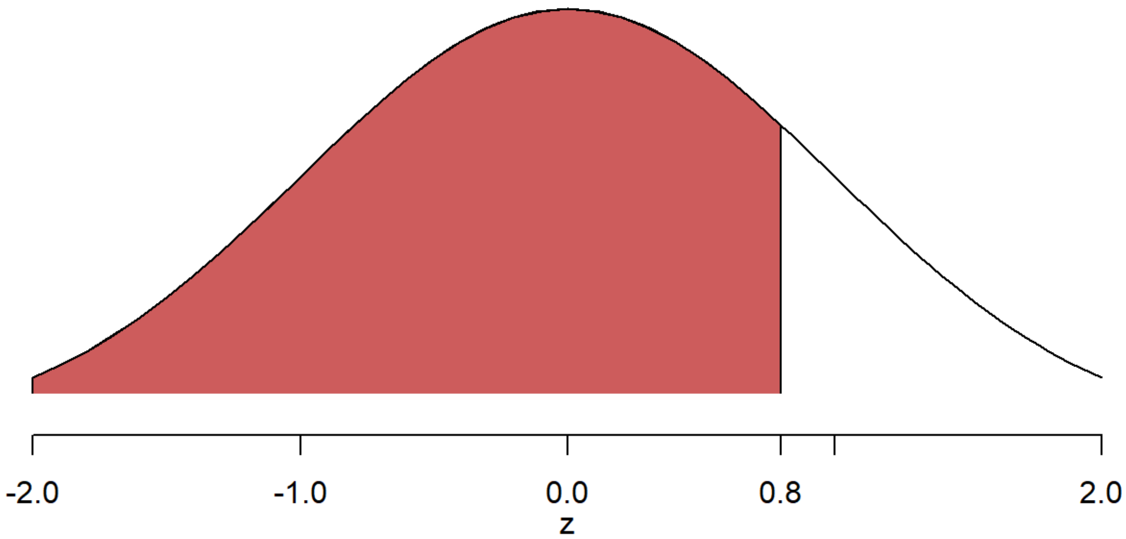


Why might we do this?

Area under a Standard Normal Curve

Example

Suppose we want to find out the area up to 0.8.



Method1: Integration

Mathematically, we could use integration:

$$\text{area} = \int_{-\infty}^{0.8} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\, dy$$

But this does not have a closed form!

Method2: Normal Tables (old school!)

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

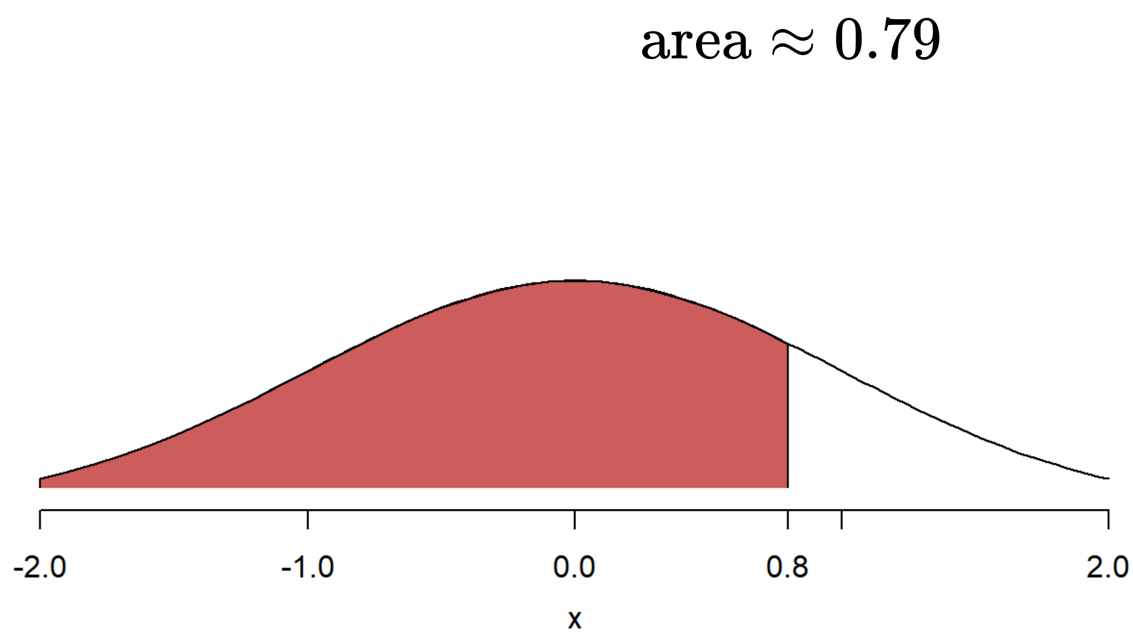
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
[illegible]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

Method3: Use R

- The `pnorm` command works out the lower tail area.
- The `pnorm (x,lower.tail=F)` works out the right tail area.

> **!** Remember
>
> Always sketch the Normal curve and the relevant area … and **then** use R!

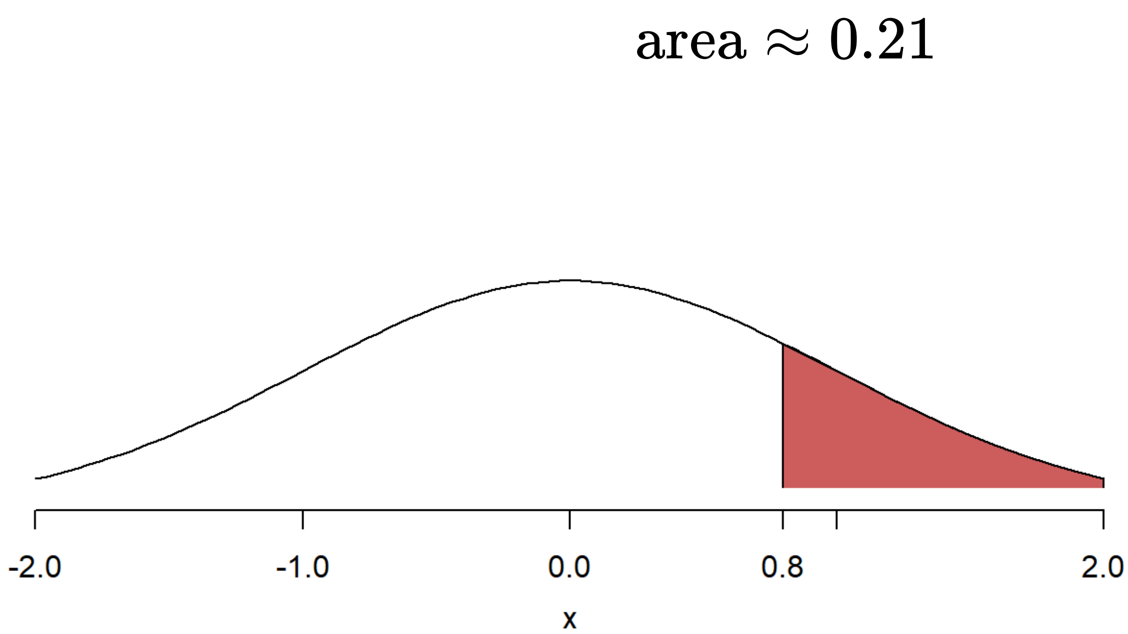
Lower tail




```r
pnorm(0.8)
```

```
## [1] 0.7881446
```

Upper tail

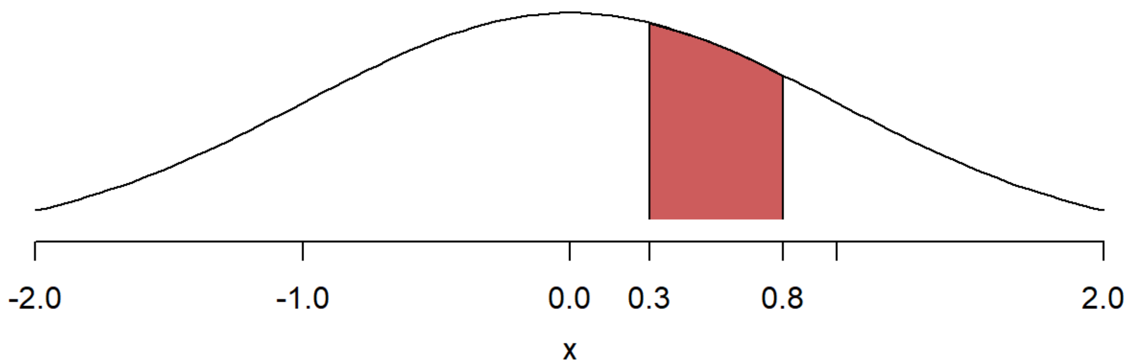



```
pnorm(0.8, lower.tail = F)
```

```
## [1] 0.2118554
```

Interval

$$\text{area} = P(Z < 0.8) - P(Z < 0.3) \approx 0.17$$

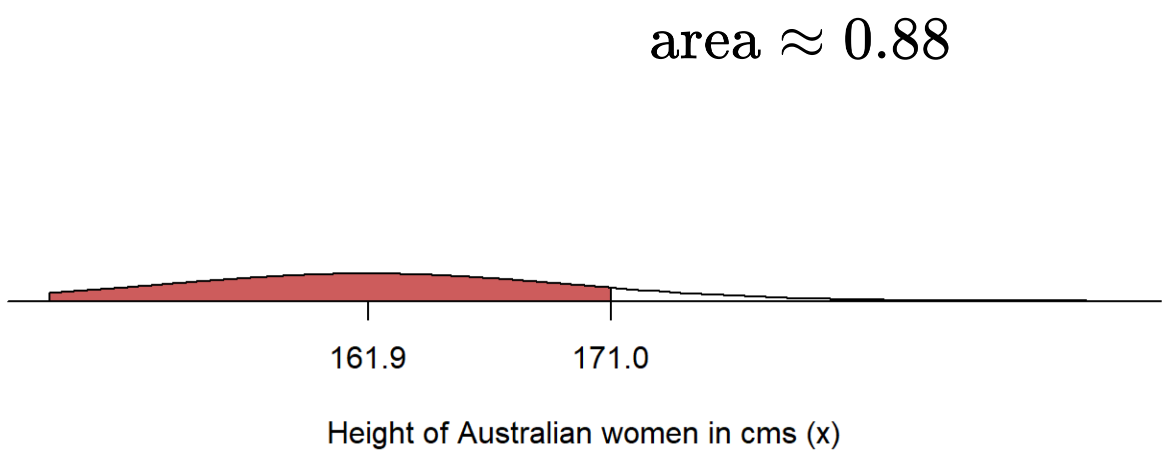



```
pnorm(0.8) - pnorm(0.3)
```

```
## [1] 0.1702332
```

Area under a General Normal Curve

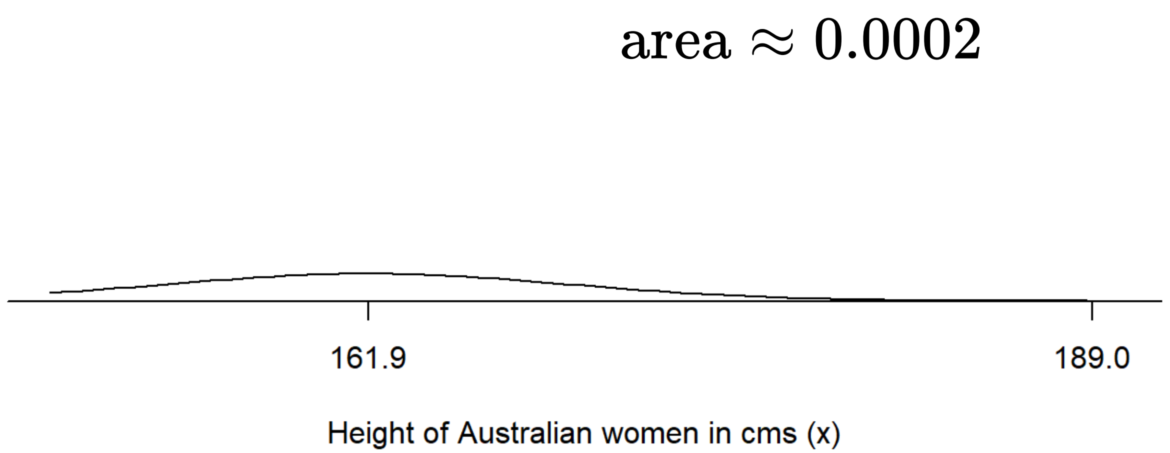
Lower tail




```
pnorm(171, 161.9, 7.7)  #pnorm(x,mean,sd)
```

```
## [1] 0.8813611
```

Upper tail




```
pnorm(189, 161.9, 7.7, lower.tail = F)
```

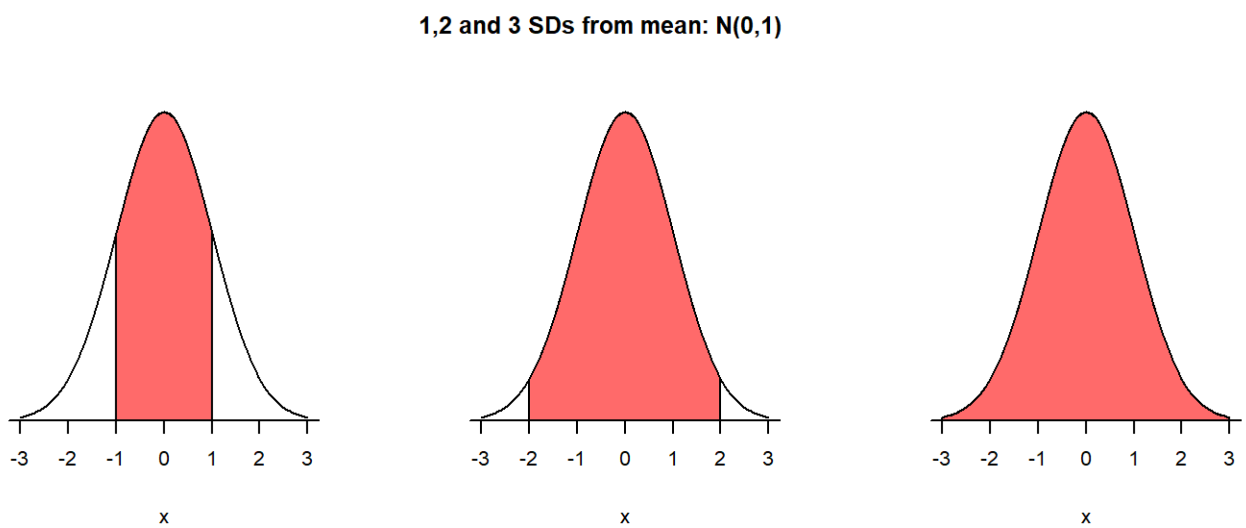
```
## [1] 0.0002161964
```

 So using the Normal curve approximation, how likely is to find an elite netball goal player in Australia?

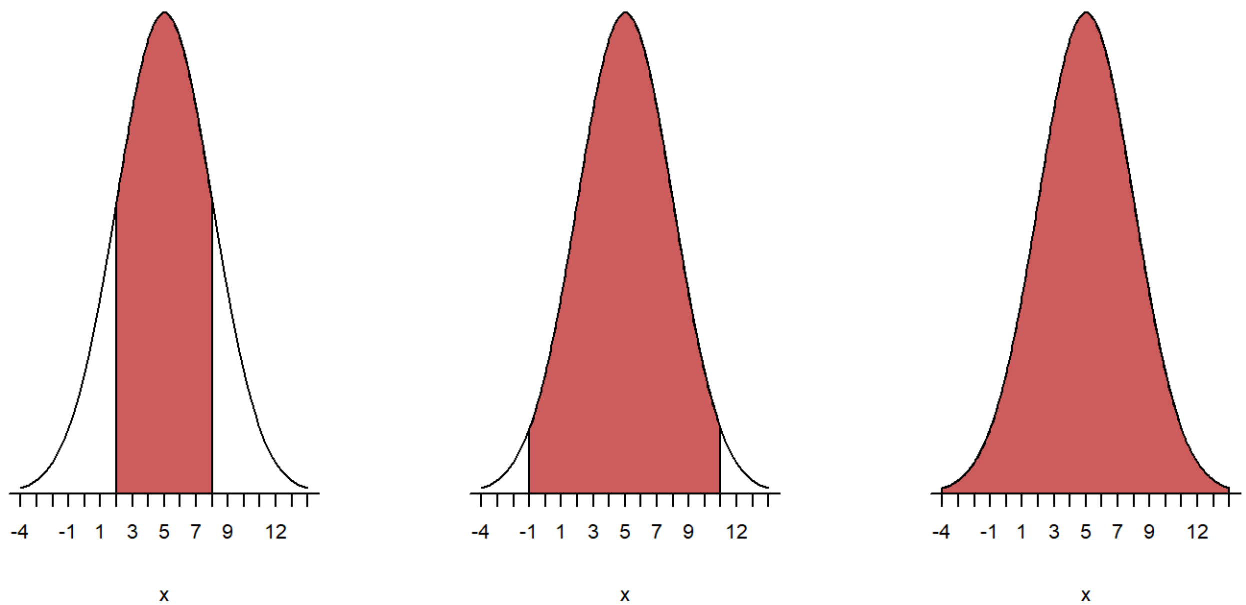
Special Properties of the Normal

1. 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%).



1,2 and 3 SDs from mean: N(5,9)



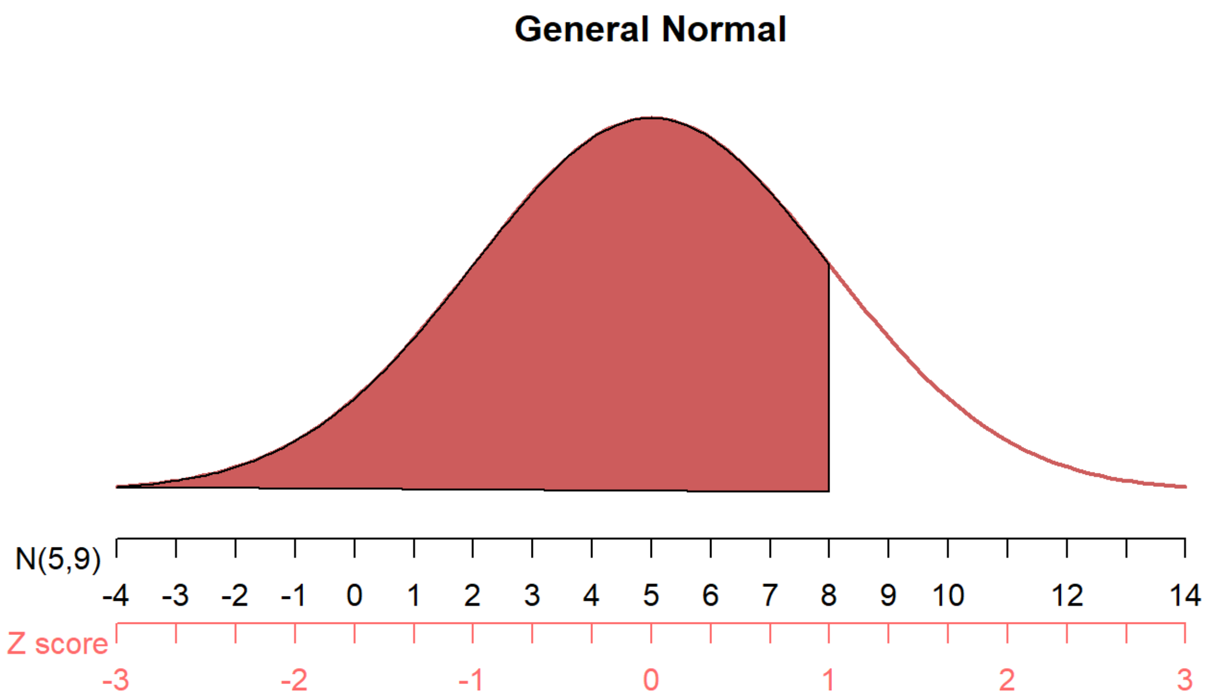
2. Any General Normal can be rescaled into the Standard Normal

Standard units

For any point on a Normal curve, the standard units (or z score) is how many standard deviations that point is above (+) or below (-) the mean.

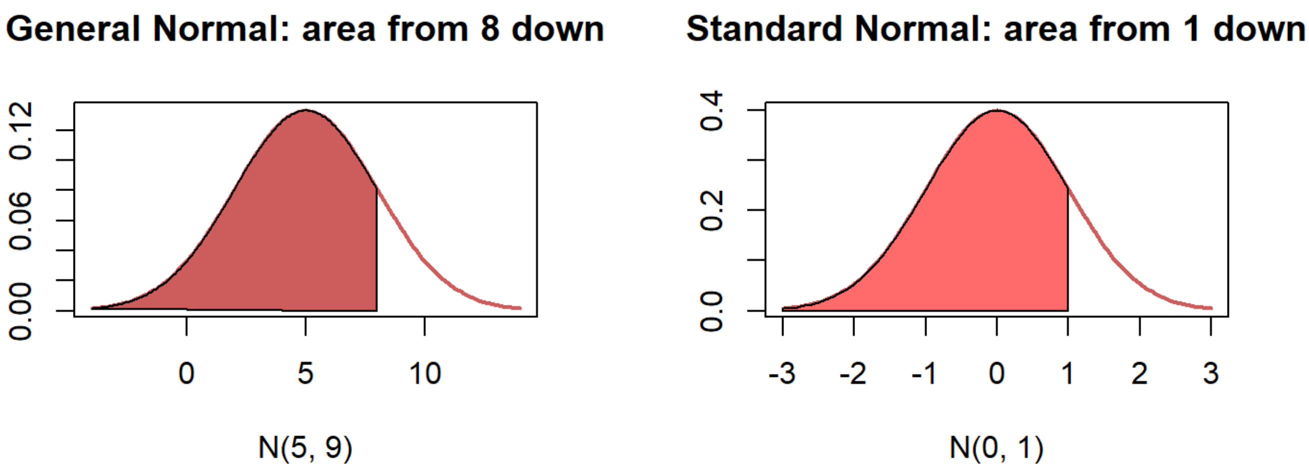
$$\text{standard units} = \frac{\text{data point - mean}}{\text{SD}}$$

Example1

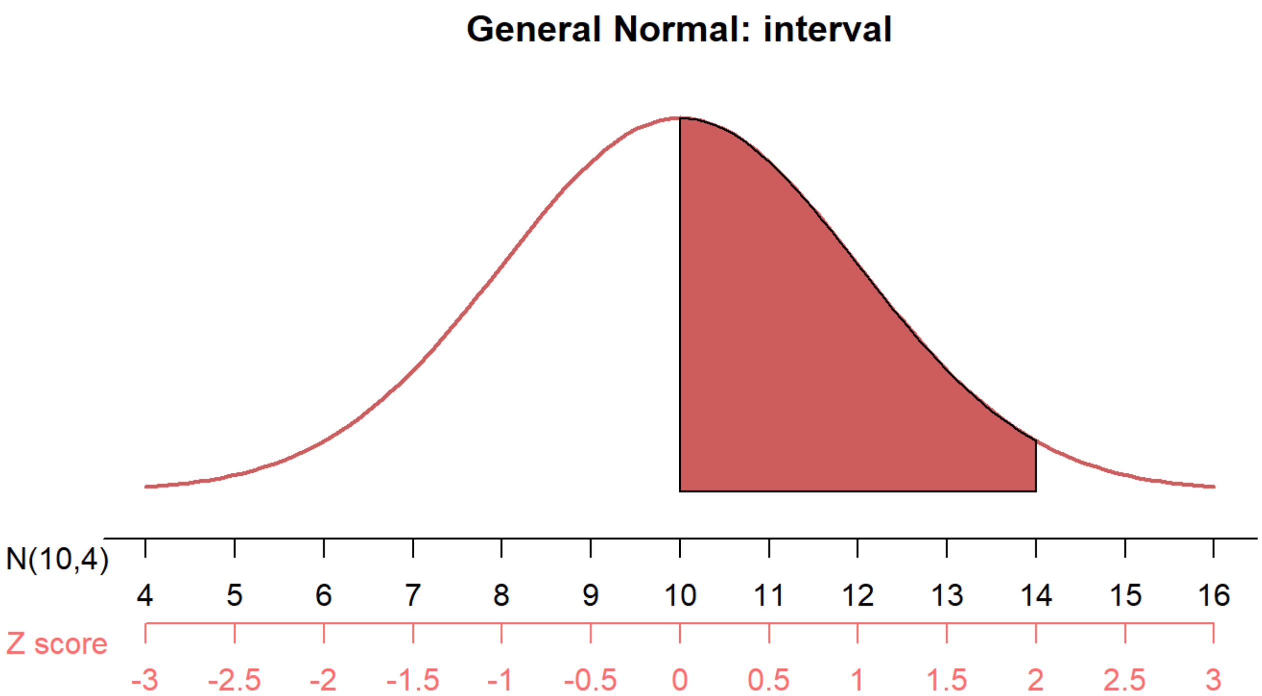


- Consider the point = 8.
- So the z score is $\frac{8-5}{3} = 1$.

The following 2 areas are equivalent.

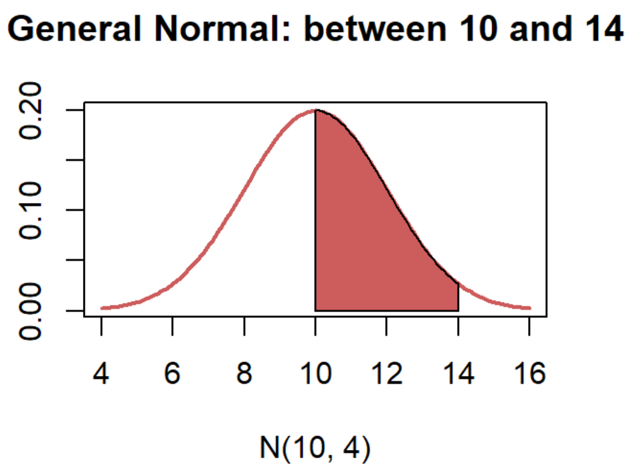


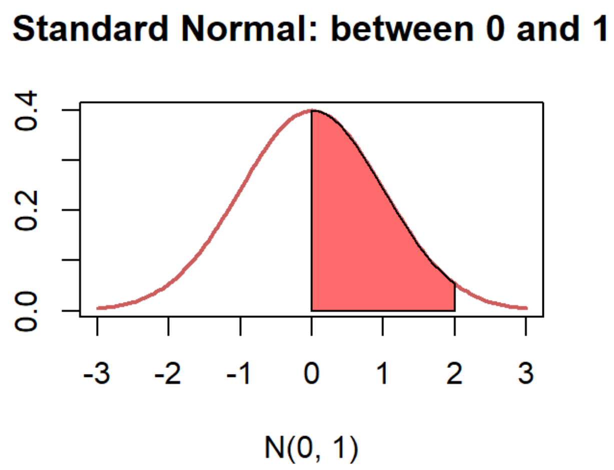
Example2



- 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 equivalent.





Modelling using the Normal Curve

Limitations of using any model

All models are approximations. Assumptions, whether implied or clearly stated, are never exactly true. All models are wrong, but some models are useful. So the question you need to ask is not “Is the model true?” (it never is) but “Is the model good enough for this particular application?”

Box, G., Luceno, A. & del Carmen Paniagua-Quiñones, M. (2009) *Statistical Control By Monitoring and Adjustment*, p61.

How do we know when to use the Normal curve?

1. Does the histogram look Normal?

Is the histogram basically balanced without long tails or many outliers?

2. Do the proportions look right?

```
sum(data2[data2 > mean(data2) - sd(data2) & data2 < mean(data2) + sd(data2)])/sum(data2)
```

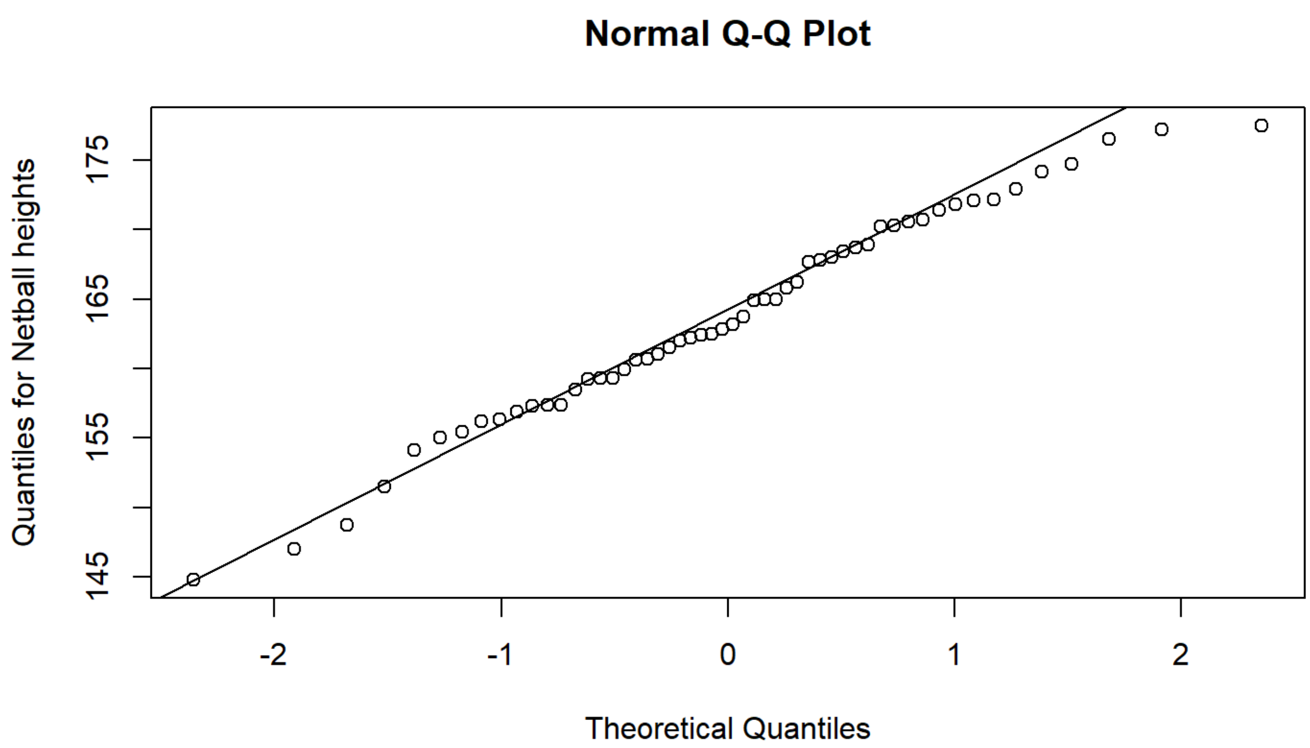
```
## [1] 0.6859907
```

```
sum(data2[data2 > mean(data2) - 2 * sd(data2) & data2 < mean(data2) + 2 * sd(data2)])/sum(data2)
```

```
## [1] 0.955145
```

3. Does the quantile-quantile (QQ) plot look like a straight line?

```
qqnorm(data$heights, ylab = "Quantiles for Netball heights")
qqline(data$heights)
```



4. Shapiro test (Ext)

```
shapiro.test(data$heights)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$heights
## W = 0.98237, p-value = 0.6067
```

Note: The Shapiro test tests the hypothesis that the Normal curve fits the data. A p-value less than 0.05 suggests that the Normal curve may not fit. So here the Normal curve appears to be an adequate fit. More of this in Hypothesis Testing.

Summary

The Normal curve naturally describes many histograms, and so can be used in modelling data. It has many useful properties, including the 68/95/99.7% rule and that any General Normal can be rescaled into a Standard Normal.

Key Words

de Moivre, Standard Normal curve, General Normal curve, standard unit (z score), diagnostics

Further Thinking