Topic6: Continuous Random Variables

Outline

Example: Australian Netball and AFL teams

In 2015 the Australian Institute of Sport ran a netball training camp for the best Australian young players playing in goals: shooters, attackers, keepers and defence. All players were over 189cm in height.

'Tall goal shooters and tall goal keepers are much more a part of the scene than when I was playing 15 years ago. We recognise that in Australia the game of netball is changing, we want to remain competitive and maintain a competitive advantage."
(Former Australian netball team member and AIS Centre of Excellence coach Jenny Borlase) ▸ ABC News

What is the probability of finding an Australian woman of 'goal player' height or taller?

If X represents the heights of Australian women (in cms), what is $P(X > 189)$?

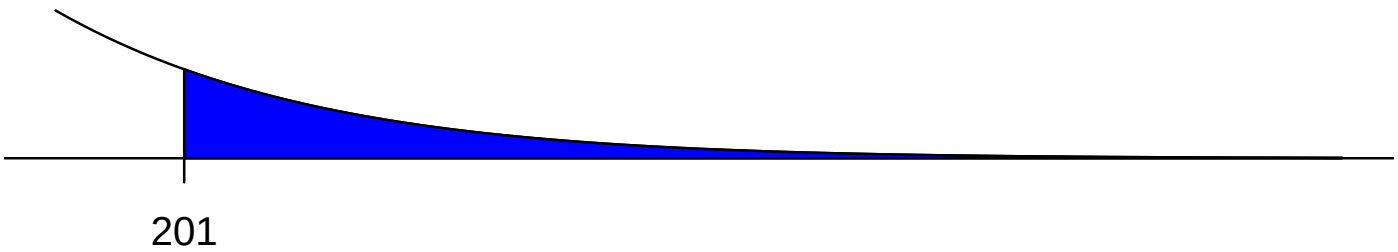
Similarly, in the Australian Football League (AFL) recruiters tend to look for tall male players.

'Data collected by Adelaide sports doctor Geoffrey Verrall and AFL sports scientist Jamie Hepner suggest genetics play a bigger part in achieving a top-level football career than skill or desire. Using AFL statistics, they noted the height of the 562 rookies selected in AFL national drafts between 2004 and 2010. They found those shorter than 180cm could just about forget trying for a place in the elite league - unless they happened to be an indigenous or Pacific Islander player, who thrive on leg speed and lightning reflexes ... Generally, white Australian males must be tall (about 188cm) to make it in the AFL.' ▸ HeraldSun

What proportion of Australian men are similar height to Sydney superstar Adam Goodes (191cm) or West Coast ruckman Nick Naitanui (201cm)?

▸ SBS

If Y represents the heights of Australian men (in cms), what is $P(Y > 191)$ or even $P(Y > 201)$?



Comparing Discrete and Continuous Distributions

There are 5 fundamental differences between discrete and continuous distributions:

	Discrete	Continuous
Values	Countable	Infinite
Plot	Histogram $P(X = x)$ probability distribution function	Smooth curve $f(x)$ probability density function (pdf)
$P(X = x)$	$0 \leq P(X = x) \leq 1 \ \ \forall x$	$P(X = x) = 0 \ \ \forall x$
Sum of Probabilities	$\sum_x P(X = x) = 1$ Area of histogram	$\int_x f(x)dx = 1$ Area under density
$F(x) = P(X \leq x)$ CDF ▸ CDF	$\sum_{y=min(x)}^{x} P(X = y)$	$\int_{-\infty}^{x} f(y)dy$

For any continuous distribution:

- there is an infinite number of possible values;
- these values may be within a fixed interval. For example, male human heights (in cm) belong to [54.6,272]. ▸ Human Heights
- each of the individual probabilities is 0, ie $P(X = x) = 0 \ \ \forall x$. This looks strange at first. However, consider that if we allocate even the smallest amount of probability to each of the infinite values, the probabilities could never sum to 1!
- the total of all the probabilites, represented by the area under the probability density function (pdf), must be 1.
- For a continuous distribution

$$P(a < X < b) = P(a \leq X \leq b)$$

This is not generally true for a discrete distribution.

Normal Distribution

Definition (Normal Distribution)

The Normal distribution models a symmetric, bell-shaped variable with 2 parameters mean μ and variance σ^2 and points of inflection at $\mu \pm \sigma$. We say the variable $X \sim N(\mu, \sigma^2)$.

The probability density function (pdf) is:

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

The cumulative distribution function (CDF) is

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)dy$$

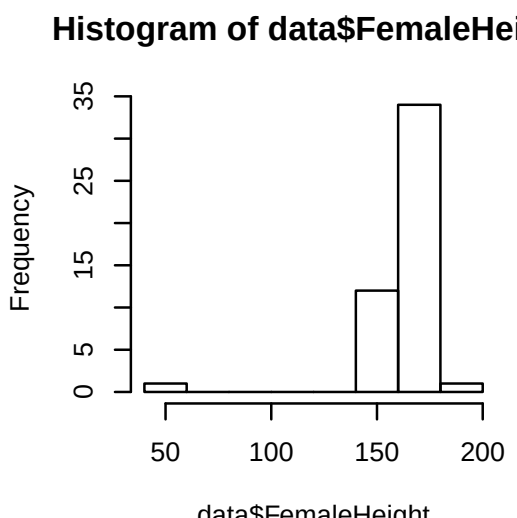
The Normal Distribution is very important because it can approximate many natural phenomemon, like annual rainful (sometimes skewed), humidity, evapotranspiration, heights/weights/length of animals, intelligence, and measurement errors.

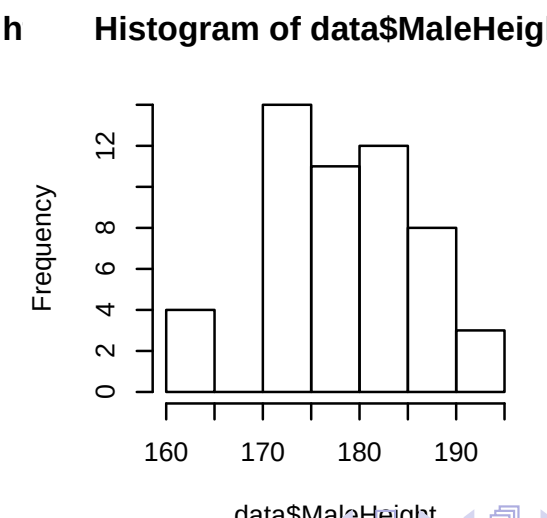
It can also approximate sums of random variables, via the Central Limit Theorem (Topic 7).

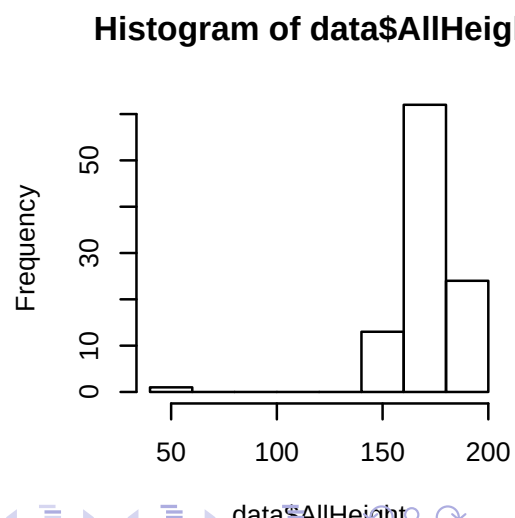
Does the Normal Distribution approximate the distribution of human heights?

Data

```
## data <- read.csv("SchoolCensusHeights.csv")
names(data)

## [1] "FemaleHeight" "MaleHeight"   "AllHeight"

par(mfrow=c(1,3))
hist(data$FemaleHeight)
hist(data$MaleHeight)
hist(data$AllHeight)
```

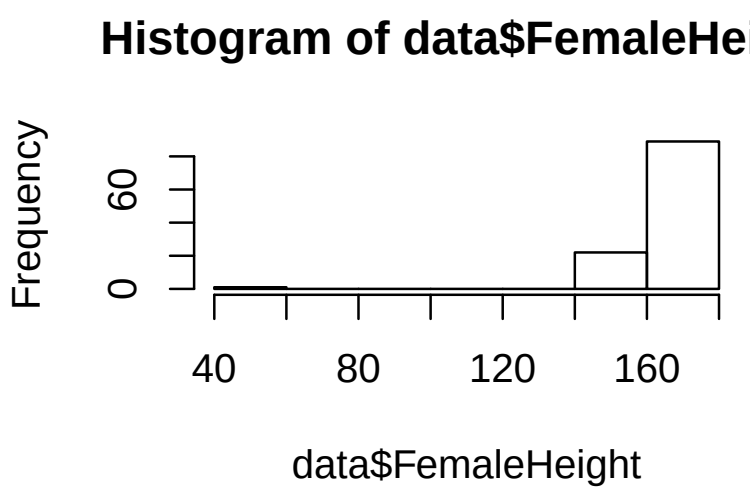


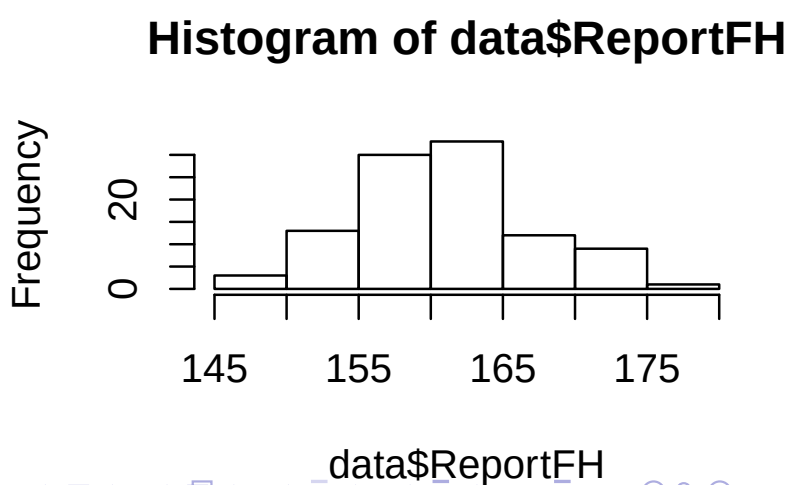




Data

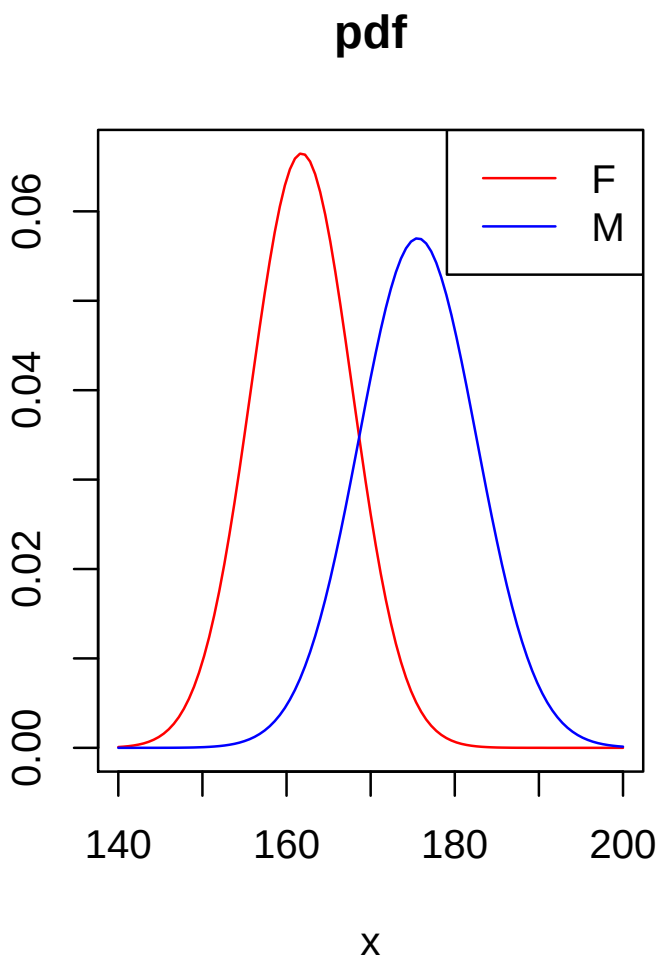
```
## data <- read.csv("DavisHeights.csv")
names(data)

## [1] "FemaleWeight" "MaleWeight"   "FemaleHeight" "MaleHe
## [5] "ReportFW"     "ReportMW"     "ReportFH"     "Report

par(mfrow=c(1,2))
hist(data$FemaleHeight)
hist(data$ReportFH)
```

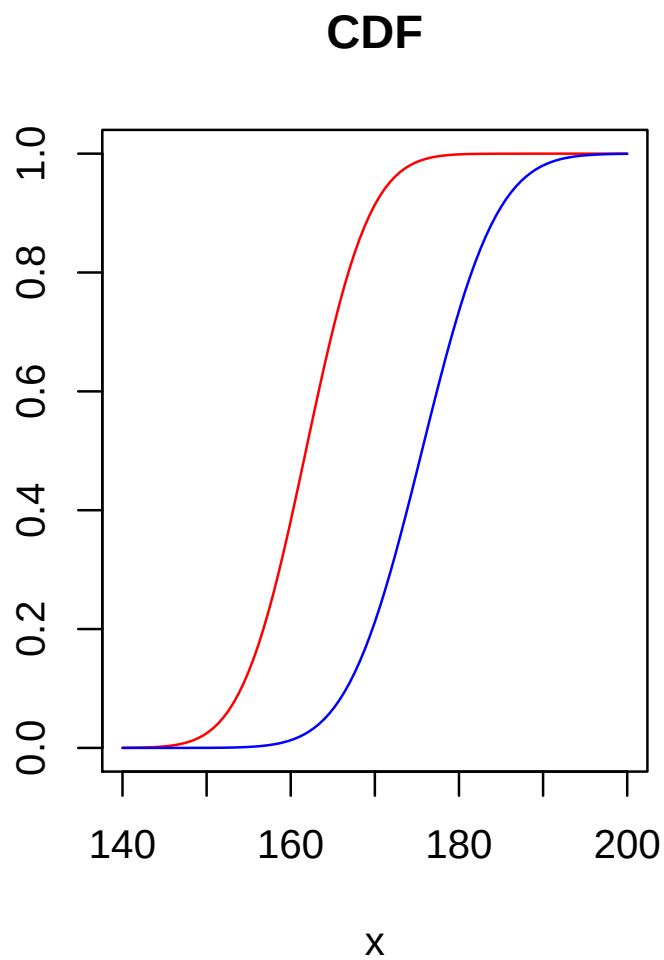




Modelling Australian women's heights by $X \sim N(161.8, 6^2)$ and men by $Y \sim N(175.6, 7^2)$, we get

▸ ABSData

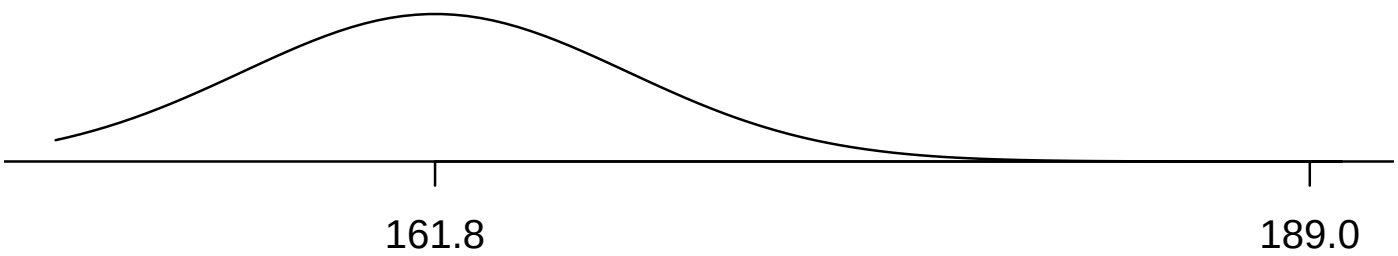




Normal Probabilities

What is the probability of finding an Australian woman of 'goal player' height or taller?

If $X \sim N(161.8, 6^2)$, what is $P(X > 189)$?



Height of Australian women in cms (x)

Method1: Integrate the pdf

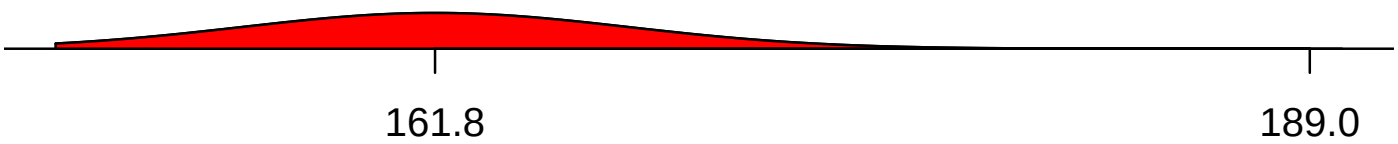
$$P(X > 189) = \int_{189}^{\infty} \frac{1}{\sqrt{2\pi(6^2)}} e^{-\frac{(y-161.8)^2}{2(6^2)}} dy$$

There is no closed form, but we could Numerical Integration.

```
f <- function(x) {dnorm(x,161.8,6)}
integrate(f,189,200)

## 2.902907e-06 with absolute error < 3.2e-20
```

Method2: Use R

```
pnorm(189,161.8,6)   #pnorm(x,mean,sd)

## [1] 0.9999971
```



Height of Australian women in cms (x)

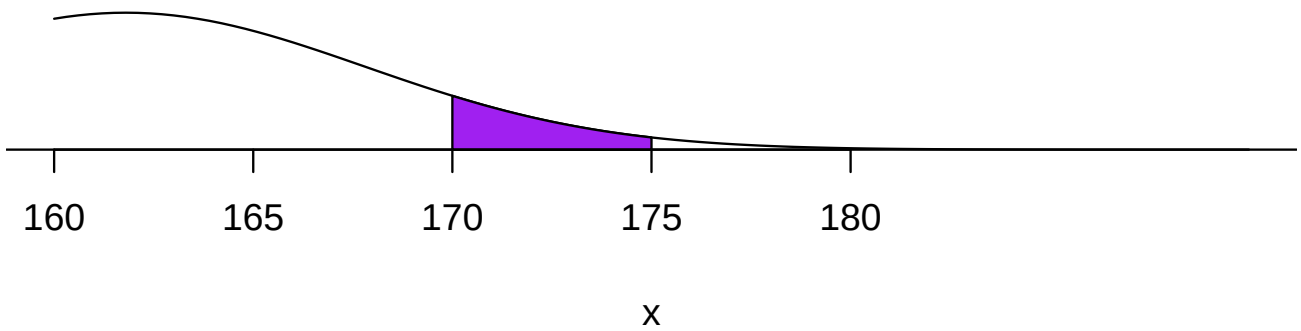
Upper Tail probabilities are found from Lower Tail probabilties:

$$P(X > 189) = 1 - P(X \leq 189) = 2.9e - 06$$

Notes on using R:

- Interval probabilities are found by subtraction:

$$P(170 < X \leq 175) = 0.07196146$$

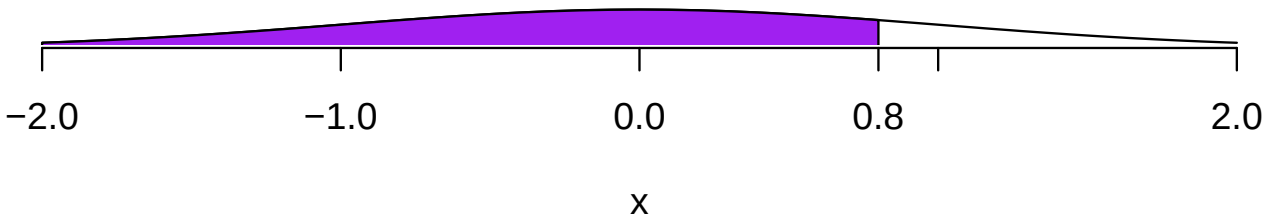



```
pnorm(175,161.8,6)-pnorm(170,161.8,6)
```

```
## [1] 0.07196146
```

- For the standard Normal $Z \sim N(0, 1)$, we can leave the mean and standard deviation unspecified.

$$P(Z \leq 0.8) = 0.7881446$$




```
pnorm(0.8,0,1)
```

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

```
pnorm(0.8)
```

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

Method3: Standardise and use the Standard Normal Tables

The Standard Normal Tables tabulate the CDF for $Z \sim N(0,1)$.

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

For example, $P(Z \leq 0.8) = 0.7881$.

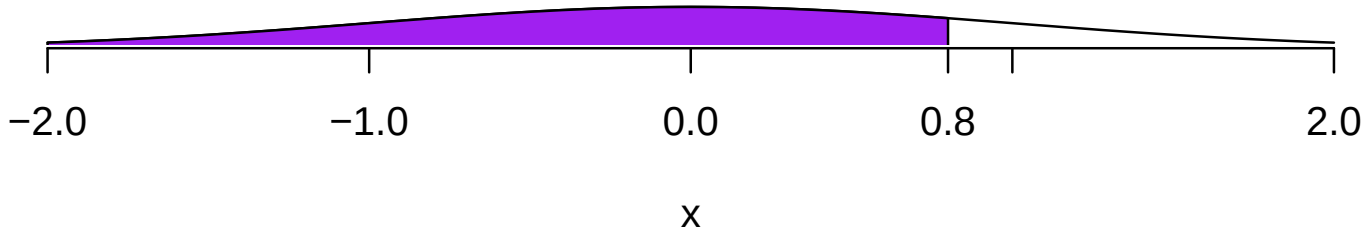


TABLE 1. **Lower tail areas of the Standard Normal distribution (CDF)** The point tabulated is $\Phi(z) = P(Z \leq z)$, where $Z \sim N(0, 1)$.

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

How do we use the Standard Normal Tables for a general Normal?

Every General Normal $X \sim N(\mu, \sigma^2)$ can be transformed into the Standard Normal $Z \sim N(0, 1)$.

Definition (Standardardising a Normal)

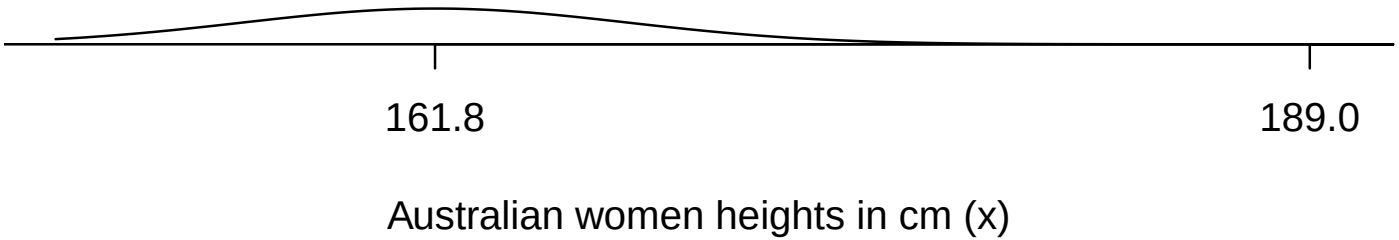
If $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$, then
$P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right)$

$$\begin{aligned} P(X > 189) &= P(\frac{X - 161.8}{6} > \frac{189 - 161.8}{6}) \\ &= P(Z > 4.533333) \\ &< 1 - 0.9986 \\ &= 0.0014 \end{aligned}$$

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
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9974
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9	
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	9	.9981
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979		
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	6	.9986

Effectively we have found that



is equivalent to

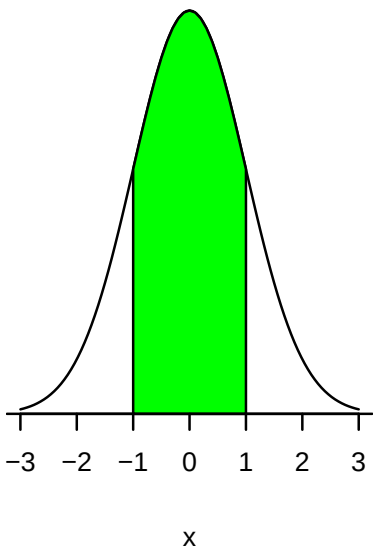

```
1-pnorm(4.533333)
```

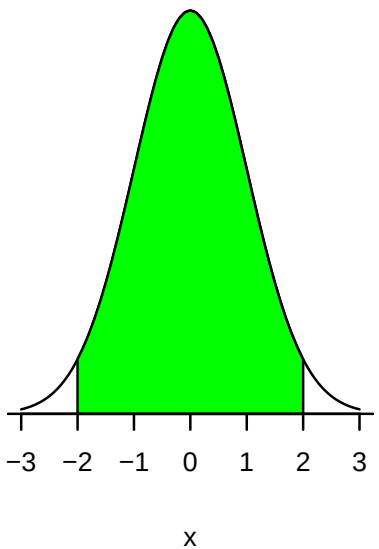
```
## [1] 2.903009e-06
```

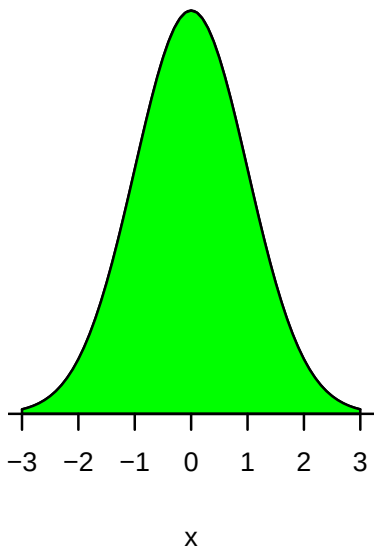
Method4: Appoximate using the Special Percentiles of the Normal Distribution

All Normal distributions satisfy the "68%-95%-99.7% Rule.

Number of sds σ from the Mean μ	% of Probability
1	68%
2	95%
3	99.7%



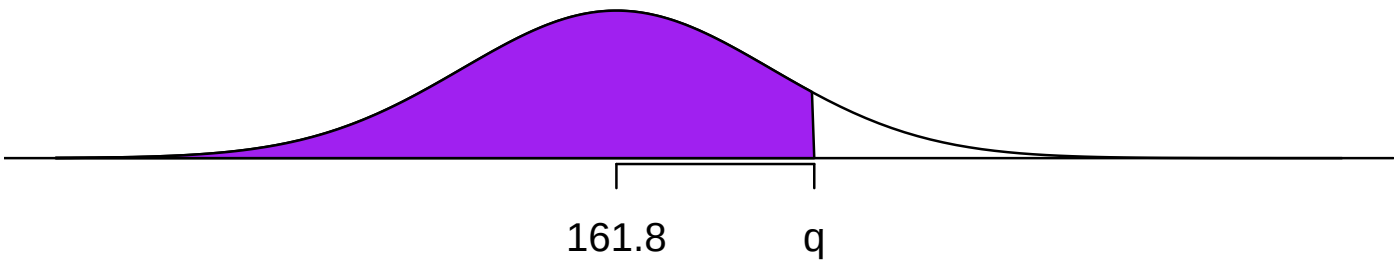




Normal Percentiles (Inverse Probabilities)

Given $X \sim N(161.8, 6^2)$, what is the 90% percentile for heights of Australian women.

We need to find q such that $P(X \leq q) = 0.9$.




```
qnorm(0.9,161.8,6)   #qnorm(%,mean,sd)

## [1] 169.4893
```

Other Continuous Distributions

We now consider 3 other continuous distributions which will be used in the Hypothesis Testing part of the course (Topic 8 onwards).

- Student T (T tests)
- Fisher's F (Test for equal variance and ANOVA in STAT2)
- Chi-Squared χ^2 (Goodness of Fit Tests)

Student T

Definition (Student T Distribution)

The Student T distribution is symmetric and bell-shaped with thicker tails than a Normal.

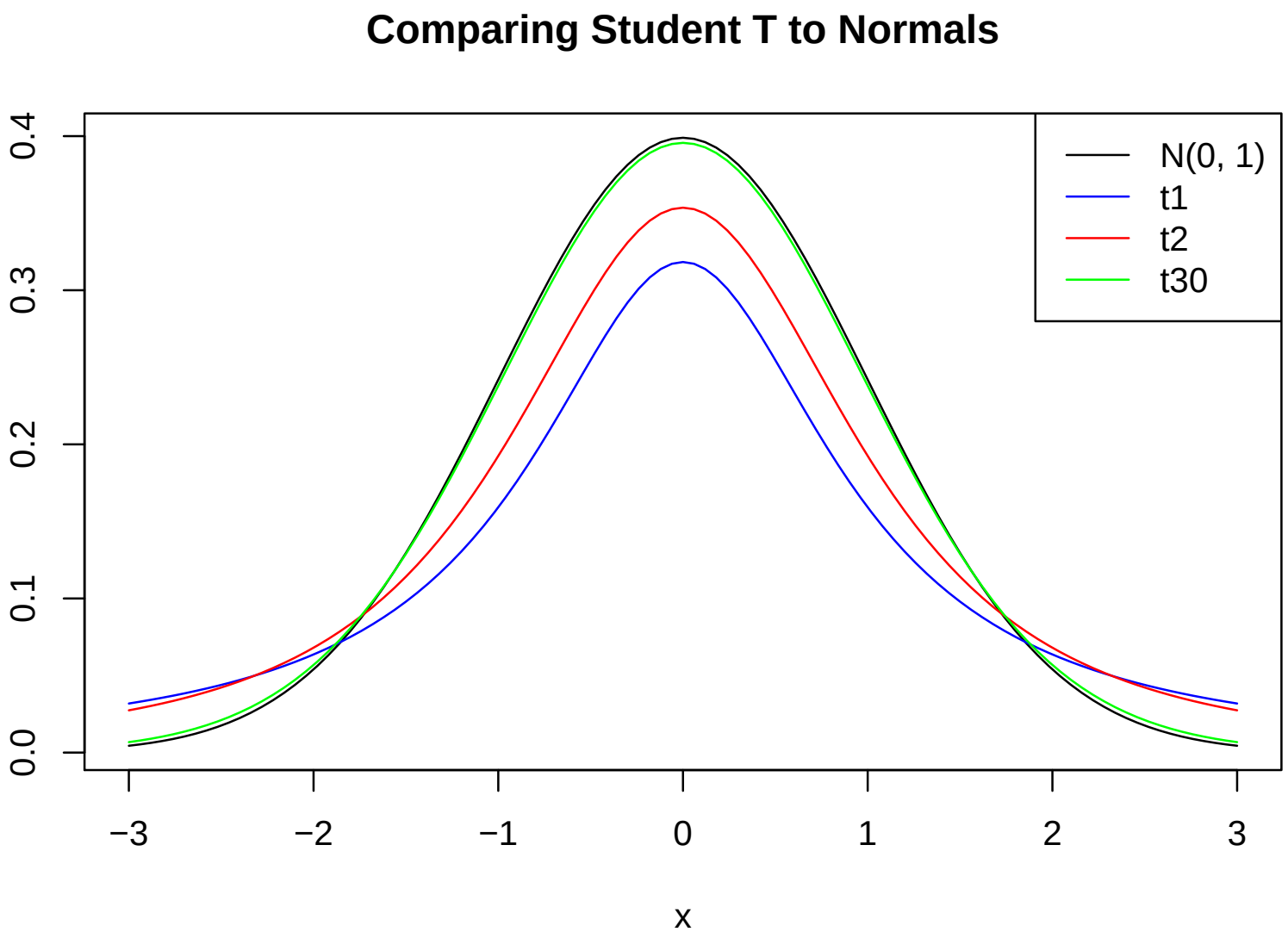
We say the variable $X \sim t_n$, with n degrees of freedom.

The pdf is:

$$f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})}(1 + \frac{x^2}{n})^{-\frac{n+1}{2}} \qquad \text{for } x \in (-\infty, \infty)$$

▸ Compare Normal

The mean is 0 and variance is $\frac{n}{n-2}$ for $n > 2$.


Comparing Student T to Normals
N(0, 1)
t1
t2
t30
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x

Chi-Squared distribution

Definition (Chi-Squared distribution)

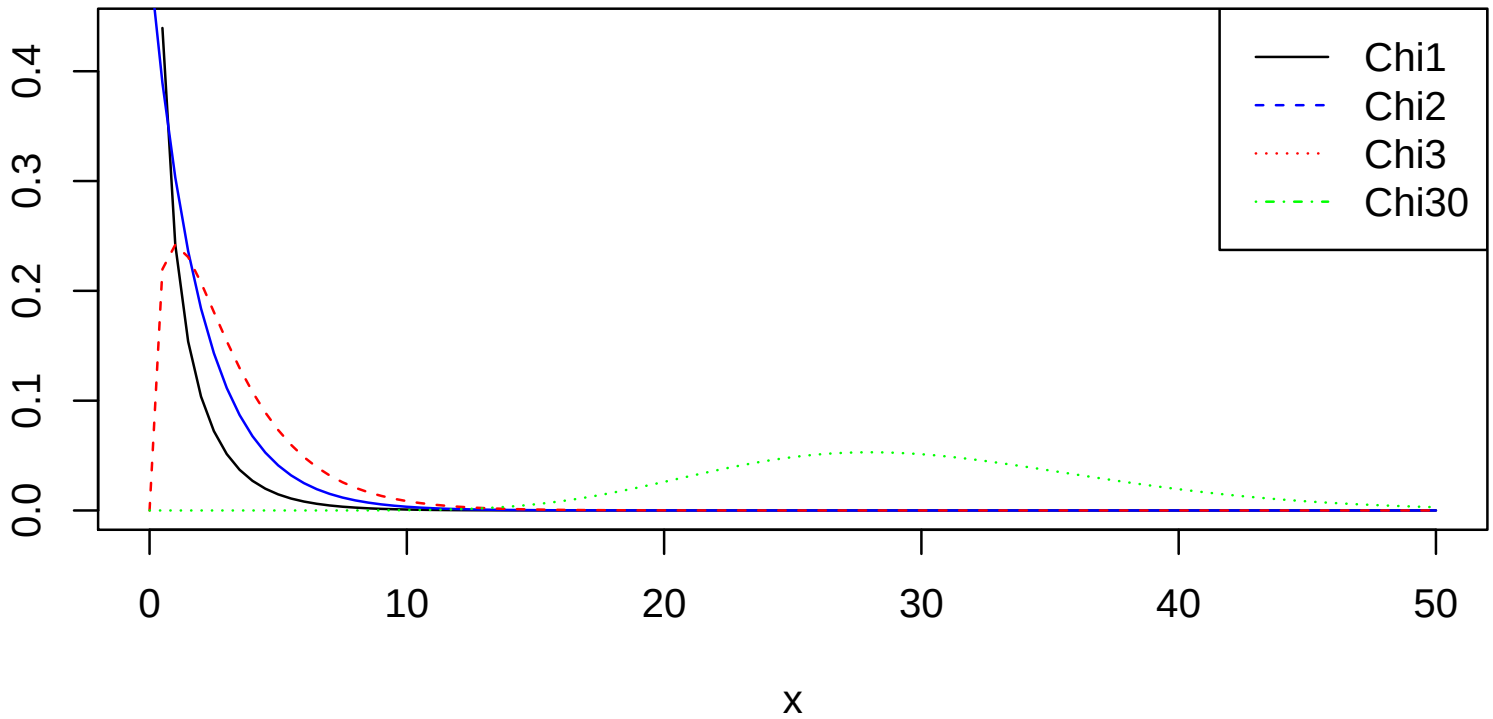
The Chi-Squared distribution is the sum of squared independent Standard Normal random variables. It can only take positive values and typically right skewed.

We say the variable $X \sim \chi^2_n$, with n degrees of freedom.

The pdf is:

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \qquad \text{for } x \in (0, \infty)$$

The mean is n and variance is $2n$.


Chi1
Chi2
Chi3
Chi30
0.0
0.1
0.2
0.3
0.4
0
10
20
30
40
50
x

Fisher's F

Definition (Fisher's F Distribution)

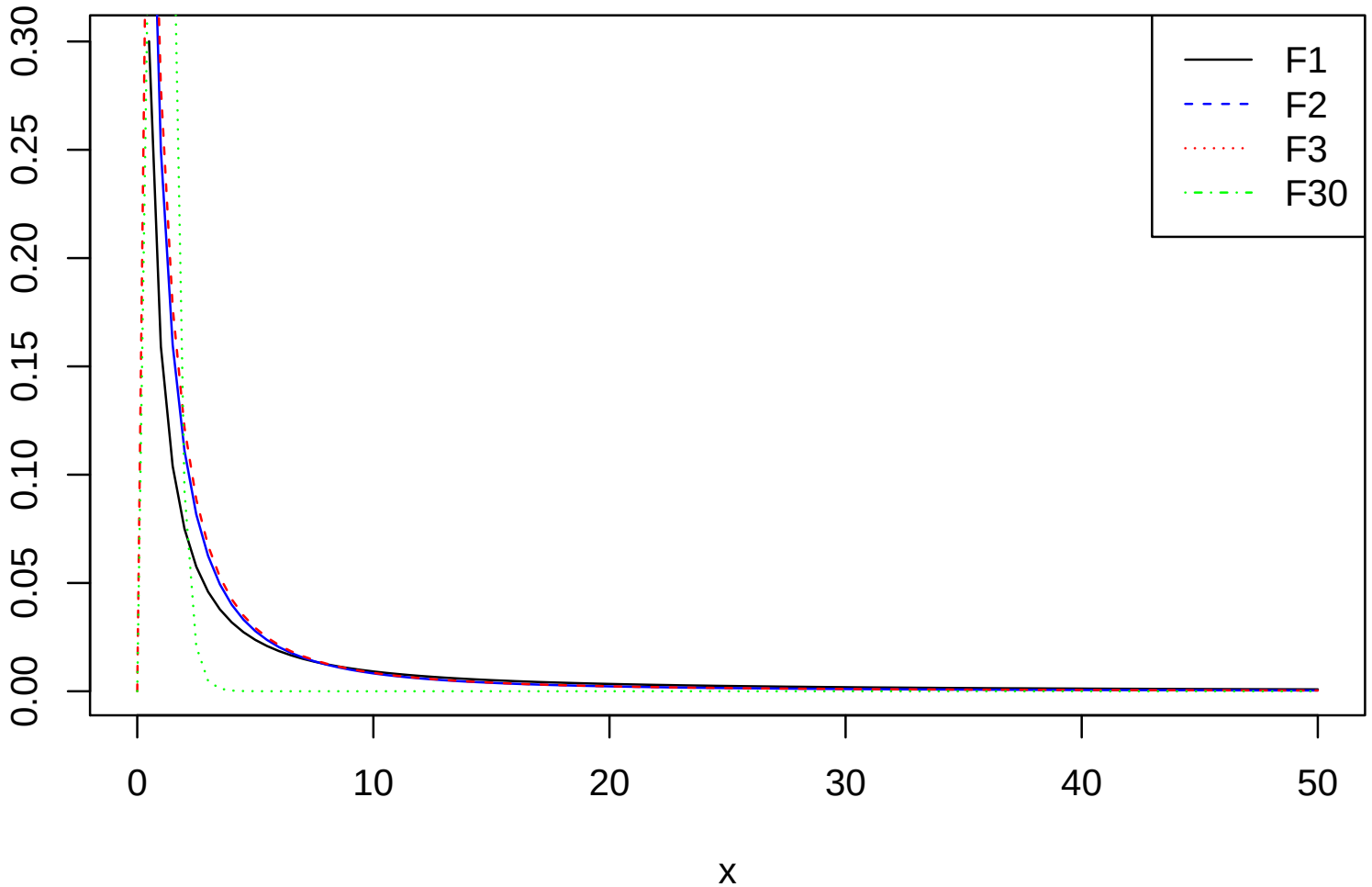
The Fisher's F distribution is the scaled ratio of 2 χ^2 variables, with m and n degrees of freedom.

We say the variable $X \sim F_{m,n}$.

The pdf is:

$$f(x) = \frac{\Gamma(\frac{m+n}{2}) m^{\frac{m}{2}} n^{\frac{n}{2}}}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})} \frac{x^{\frac{m}{2}-1}}{(n+mx)^{\frac{m+n}{2}}} \quad \text{for } x \in (0, \infty)$$

The mean is $\frac{n}{n-2}$ for $n > 2$ and variance is $\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ for $n > 4$.


F1
F2
F3
F30
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0
10
20
30
40
50
x

Using commands in R:
$P(X \leq 2)$, where $X \sim t_4$.

```
pt(2,4)
```

```
## [1] 0.9419417
```

$P(X \geq 3)$, where $X \sim \chi^2_4$.

```
1-pchisq(3,4)
```

```
## [1] 0.5578254
```

$P(1 < X \leq 3)$, where $X \sim F(2,4)$.

```
pf(3,2,4)-pf(1,2,4)
```

```
## [1] 0.2844444
```

Summary Examples

Have a go

Dharshani Sivalingam is the tallest netball player in the world. What is the probability of finding an Australian woman of Dharshani's height?

▸ Dharshani

```
#Check your answer
d=(208.3 - 161.8)/6
pnorm(d)

## [1] 1
```

Have a go

Madison Robinson is the shortest Australian International player. What percentage of Australian women are between Madison and Dharshani's heights?

▸ Madison

```
#Check your answer
m=(168 - 161.8)/6
pnorm(d)-pnorm(m)

## [1] 0.150724
```

Have a go

If 60% of Australian women are below a certain height, what is that height?

```
#Check your answer
qnorm(0.6,161.8,6)

## [1] 163.3201
```

Have a go

If $X \sim N(8, 16)$ find $P(X \leq 20)$.

```
#Check your answer
pnorm(20,8,4)

## [1] 0.9986501

pnorm(3)

## [1] 0.9986501
```

Have a go

Given Australian male heights can be approximated by $X \sim N(175.6, 7^2)$, what proportion of Australian man are similar height to Sydney superstar Adam Goodes (191cm) or West Coastruckman Nick Naitanui (201cm)?

```
#Check your answer
a = (191-175.6)/7
1- pnorm(a)

## [1] 0.01390345

n = (201-175.6)/7
1- pnorm(n)

## [1] 0.000142497
```