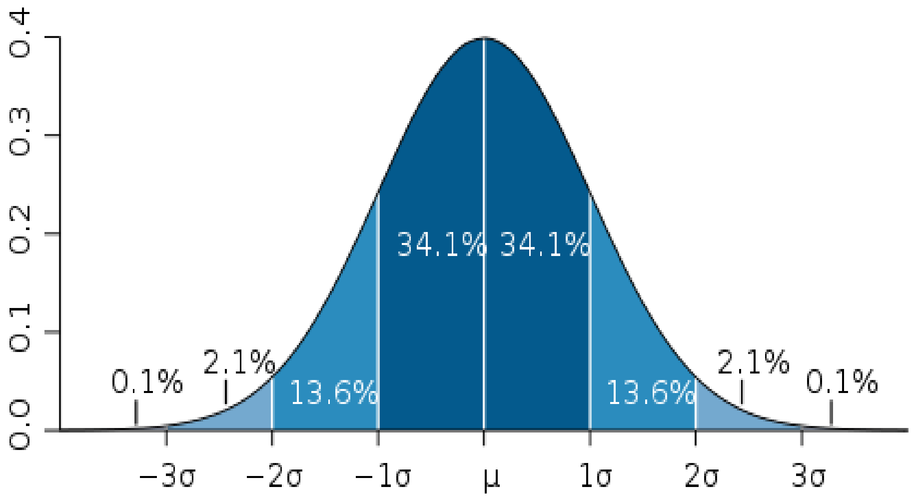
Lecture 2

Statistical inference: the normal curve and confidence intervals

Probability distributions

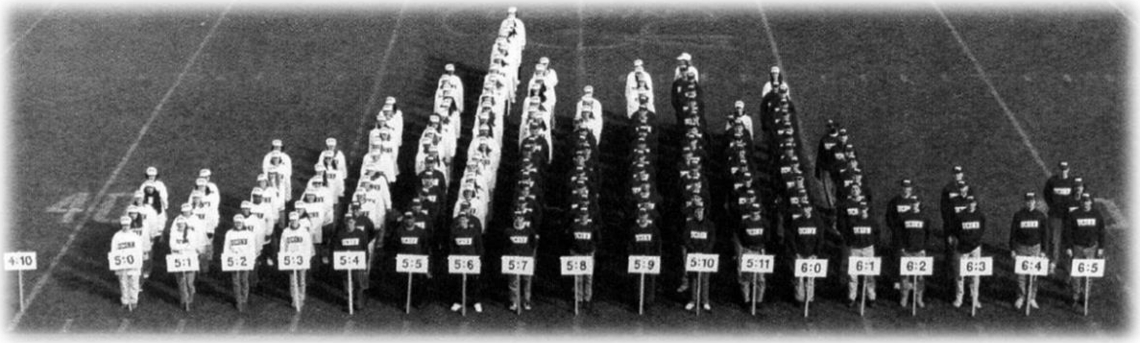
- We are now familiar with descriptive statistics; but statistical methods are mostly used for *prediction*
  - i.e. we collect samples mostly to predict characteristics of whole populations
- Extrapolation from sample to population relies on *probability distributions:*
  - a model or theory of how a variable ‘behaves’, e.g. its distribution around a mean
- In the following, we introduce the concept and uses of the Gaussian distribution (the ‘normal’ or ‘bell curve’)



Reasons for using the normal distribution

- Many characteristics of populations look ‘bell-shaped’

- Biological, social etc. traits are often bell-shaped



The normal distribution

- The *normal distribution* is an equation that produces a bell-shaped curve; its main features are:
  - mean value is the most likely value (= peak)
  - Probability of value decreases with distance to mean
  - sum of all probabilities is 100% (=the whole sample)

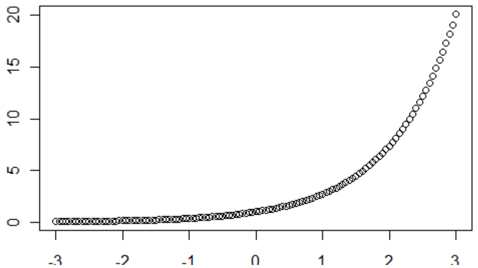
peak, symmetrical shape

- What kind of curve/distribution produces a bell-shaped curve?

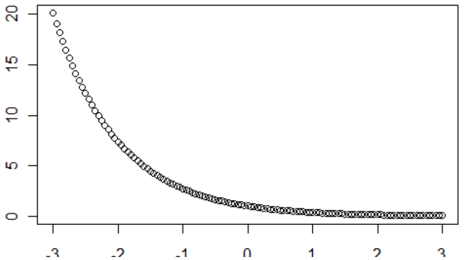
- Let's try some exponential curves
  - i.e. curves where $y = e^{f(x)}$

Y equals e to the power f(x)

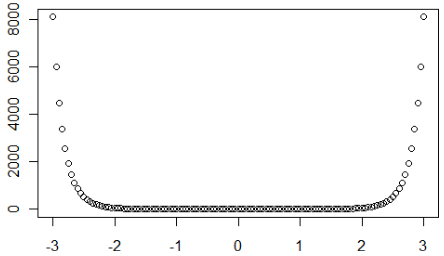
```
x <- seq(-3,3,0.05)
plot(exp(x) ~ x, type=“l”)
```



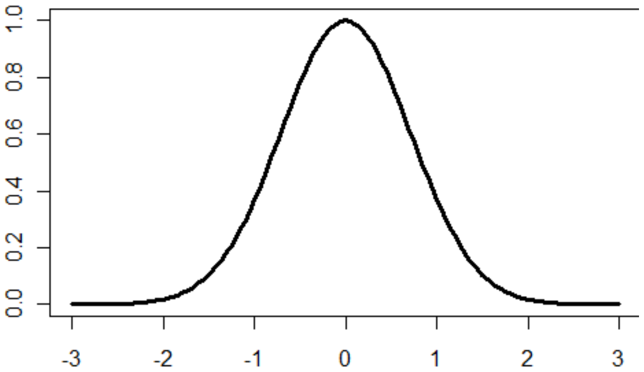

```
plot(exp(-x) ~ x, type=“l”)
```




```
plot(exp(x^2) ~ x, type=“l")
```




```
plot(exp(-x^2) ~ x, type=“l")
# that works!
```



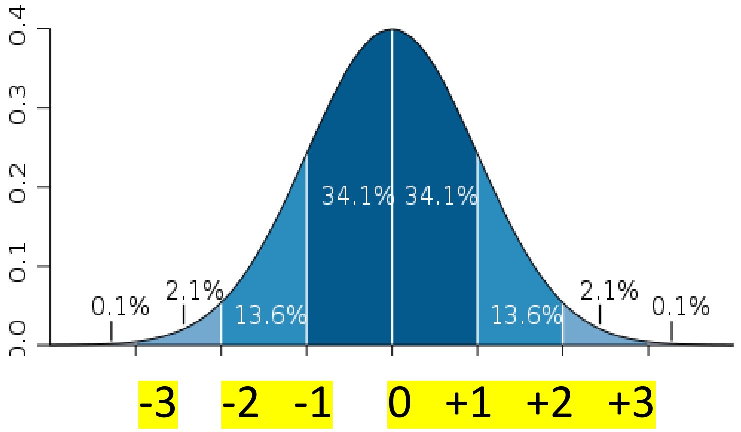
The normal equation

- The equation $y = e^{-x2}$ would work and produce a bell-shaped distribution

- The normal curve is a version of our curve:

$$N(0,1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

Mean, sd



Features:
- bell-shaped
- mean=0
- sd=1
- sum of frequencies (area under curve)=1=100%

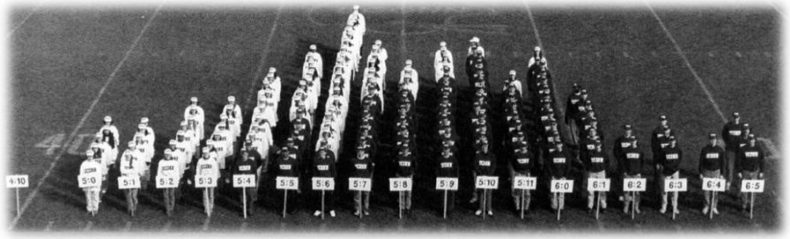
- Statisticians have analytically extracted probabilities and intervals from normal curve
  - probability of being over +3 sd from mean: 0.1%

Standardisation: everything is 'normal'

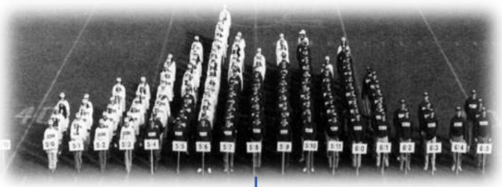
- Real traits rarely have mean=0 and standard deviation=1

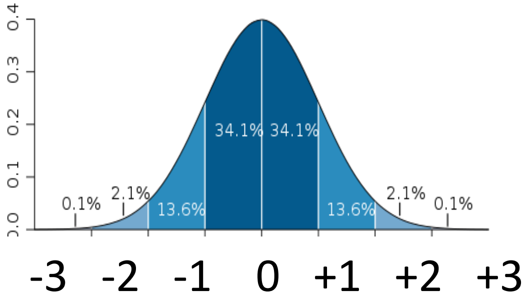
- That is not a problem: we can *standardise* variables so that *everything you measure* has mean=0 and sd=1

- How is this done? With ***z-scores***



Calculating z-scores:





If mean height is 180cm and sd=10cm:

- 1) *Subtract mean* value from each case
  - mean (μ, mu) becomes 0:
  - mean case (180cm) now measures 0
  - a 170cm-tall person now measures 170-180 =-10cm(=residual)

- 2) Divide all residuals (case value minus mean) by standard deviation
  - if sd (σ, sigma) is 10cm and mean is 180cm:
  - person measuring 170cm deviates -10cm/10cm= -1 standard deviation below the mean

Z-score

X, mean

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

sd

-1 = 170-180 / 10

- *z-score (=standardised residual) is therefore a sample-specific measure of a quantity*

2.3 = x-180/10
x= 203

Exercises:

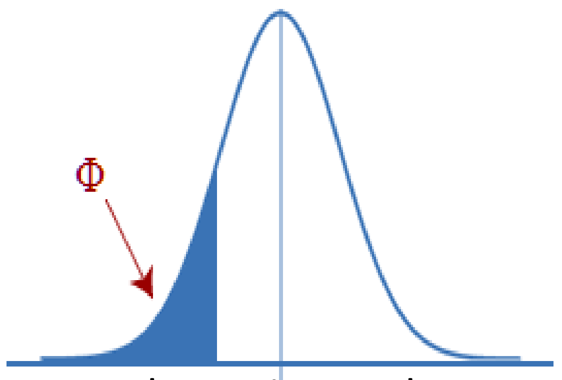
a) In this example, if a man is z=2.3, how tall is he?

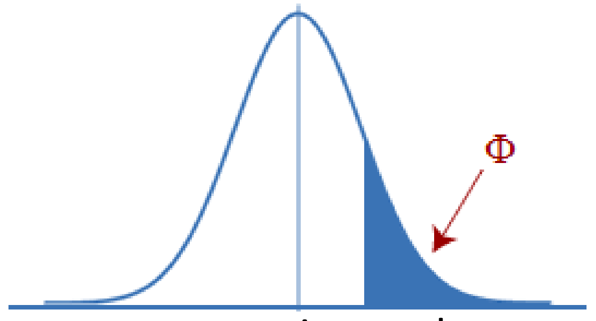
b) What's 162cm in z-scores?

x = 162-180/10
x = -1.8

Intervals and cumulative probability

- We are more interested in *intervals* of the normal curve than point values
- Why? What does it mean to ask 'what is the probability of being a millionaire in the UK?'
- It doesn't mean the probability of having *exactly* £1 million (=a point):
    - a millionaire is someone with *£1 million or over* (=an interval)
- *Cumulative probability* is the probability of an interval of values





Depends on the direction not mean

Estimating cumulative probability

Only lower intervals

- Command *pnorm(test value, mean, sd)* calculates ***cumulative*** probability *from left to right*, i.e. from -∞ to value x (the blue area) Neg infinite value up to the test value

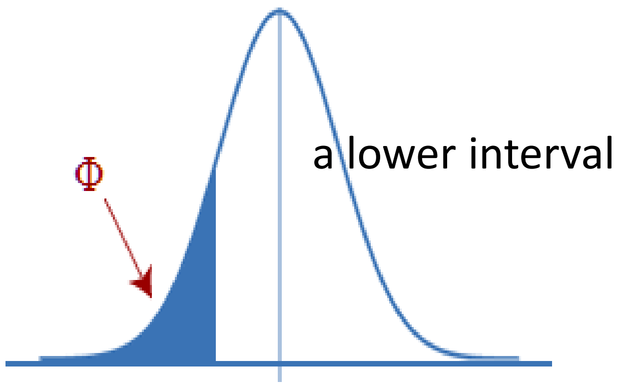
- Example: if
  - Test value = 170cm
  - mean = 180cm
  - sd =10cm

- then probability of being 170cm (=shorter than 170cm) is:

```
> pnorm(170,180,10)
[1] 0.1586553
```

=15.9%

=probability of being 1sd below mean



Upper intervals

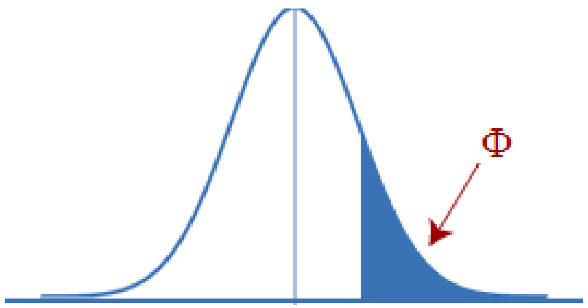
- We can use *pnorm* to estimate upper intervals too

1-pnorm to get the upper

Or

By symmetrical feature

an upper interval



Exercise:

a) If mean = 180cm and sd= 10cm, what is the probability of someone being taller than 185cm?

b) Provide answer in terms of z-score too

z=185-180/10

Probability of being ‘extreme’

- We can also calculate probability of extreme values (i.e. too large or too small)

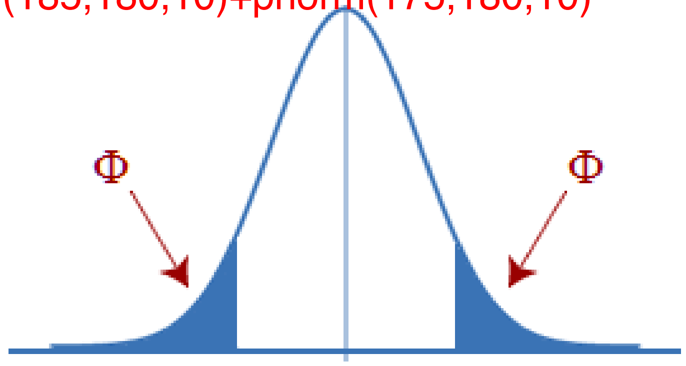
Exercise:

a) what is the probability of being shorter than 175cm OR taller than 185 cm, with N(180, 10)?

b) Provide answer in terms of z-score too

Z = 185-180 / 10

1-pnorm(185,180,10)+pnorm(175,180,10)



Now: Probability of ***not*** being an extreme case

Confidence interval

(***our most important example***)

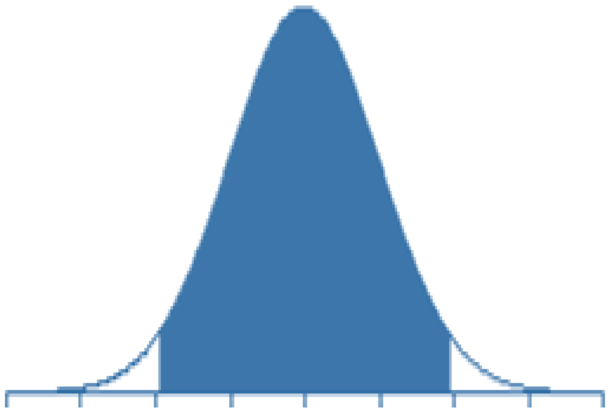
pnorm(187,180,10)-pnorm(173,180,10)

Exercise:

a) If mean = 180cm and sd= 10cm, what is the probability of someone being between 173cm and 187cm?

b) Provide answer in terms of z-score too

Z = -0.7/0.7



Statistical testing

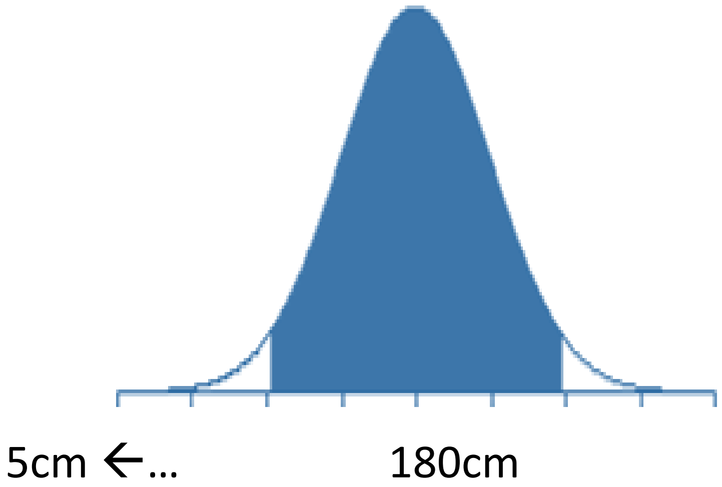
- In order to proceed to prediction and statistical testing, we need to define *confidence intervals*
- Confidence intervals are ‘acceptable’ ranges of variation, i.e. intervals including the values not differing *too much* from a population mean or expected value
- Confidence intervals are based on conventionally-defined ‘margins of error’ establishing what ‘*too much*’ means

From ‘rare’ to ‘not one of us’

- Suppose someone tells you that they’ve found 5cm-tall people on a Pacific island

- Let’s calculate the probability of a hypothetical 5cm tall human

- If our reference population has mean height=180cm and sd=10, the probability of someone being 5cm is 7.2 x 10^{-69}!




```
> pnorm(5, 180, 10)
[1] 7.163459e-69
```

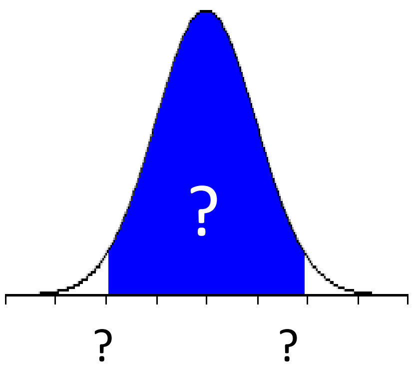
- If probability is that small, it is likely that the creatures they’ve found is not human, i.e., *they do not belong in our sample or distribution*

- (bear in mind: probability is small, but not *zero*!)

From confidence interval…

- With mean=180cm and sd=10cm, normal curve predicts that about 16% of people are shorter than 170cm; that's short, but 'human'

- But if you are 5cm tall, probability is 7.2 x 10^{-67}%; common sense says this case is too low or 'extreme' (=not human)



Question is: where, between 16% and

7.2 x 10^{-67}%, do we draw the boundary between

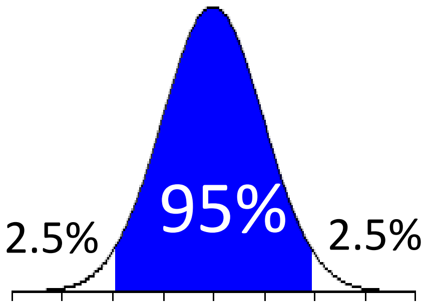
- *being rare but in the distribution (=one of us)*
- *being from another distribution? (=not ne of us)*

…to 95% confidence interval

- Answer: there is no objective limit
  - accepted limit is set conventionally:

- Most often, boundary is set at 5%
  - or less frequently, 1%
  - then, if a value is over 5% likely, i.e. within a 95% confidence interval around mean, it is accepted as part of that distribution; not ‘rare’
  - if it is less than 5% likely, it is too ‘rare’; it is defined as not in the distribution

- The conventional value of 5% defines a 95% confidence interval
  - it excludes 2.5% cases on each side, i.e. too low or too high, as not belonging in the distribution
  - It defines confidence or belief that the case belongs in the distribution



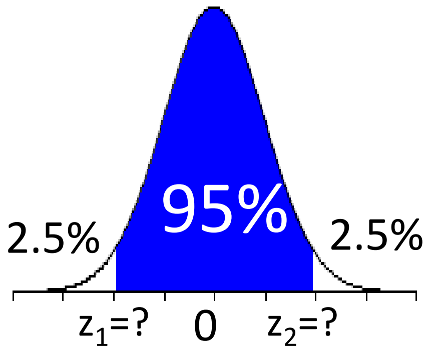
Boundaries of the 95% CI

- So if we define our CI at 95%, how much do you need to deviate from the mean to be in the 'too extreme' 5%?

Exercise:

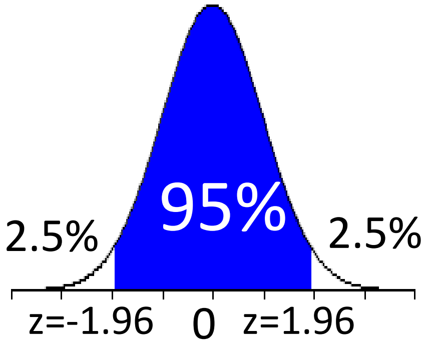
Estimate approximate lower and upper boundaries of the 95% CI using the *pnorm* function.

Present the values in z-scores and cm (assuming mean = 180cm and sd = 10cm)



Boundaries of the 95% CI

- in order to be within the 95% ‘acceptable’ values, values must be between z=-1.96 and z=1.96
  - if values less than z=-1.96 (***lower boundary***) or over z=1.96 (***the upper boundary***), they are outside confidence interval (‘too extreme’)



Tip: also try function *qnorm*:

```
> qnorm(0.025)
> qnorm(0.975)
```

To learn about qnorm (or any function):

```
> ?qnorm
```

Boundaries of the 95% CI

- Remember : if a value is **about 2 standard deviations above or below mean**, it is *outside* the 95% confidence interval

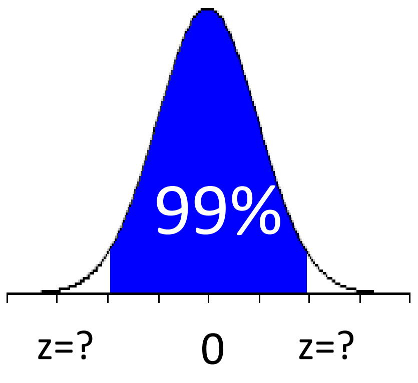
- = difference is larger than expected for a case in that sample

99% sure one event would happen, z-score should be 2.575
qnorm(0.005)

Exercise:

Estimate approximate lower and upper boundaries of the **99% CI**

Present the values in z-scores and cm (assuming mean = 180cm and sd = 10cm)



Exercises

1) Create a file with !Kung adult women only

Tips

a) use function *subset* to create a new file

b) Make a histogram of adult female weight; does the distribution look normal?

Use new file or:

```
> hist(kc$weight[kc$age > 18  & kc$sex == "woman"])
```

Brackets for additional requirements

c) How many adult females with missing weight data?

Tip: function *summary*

d) How many adult females with weight data?

e) Calculate mean and sd for adult female weight. Based on z-scores, calculate the probability of an adult woman being

i) under 40 kg

ii) over 60 kg

pnorm

mean=0, sd=1

2) Take a standardised normal distribution; what is the probability of a value being

a) Less than z=-3sd pnorm(-3,0,1)
b) greater than z=+3sd?
c) which confidence interval would those probabilities define? ? ? ? ?

- Answers to final exercises:

1)

c) 68

d) 264 – 68

2)

a) 0.001349898

b) 0.001349898

c) 99.73% CI