Lecture 2

Statistical inference: the normal curve and confidence intervals

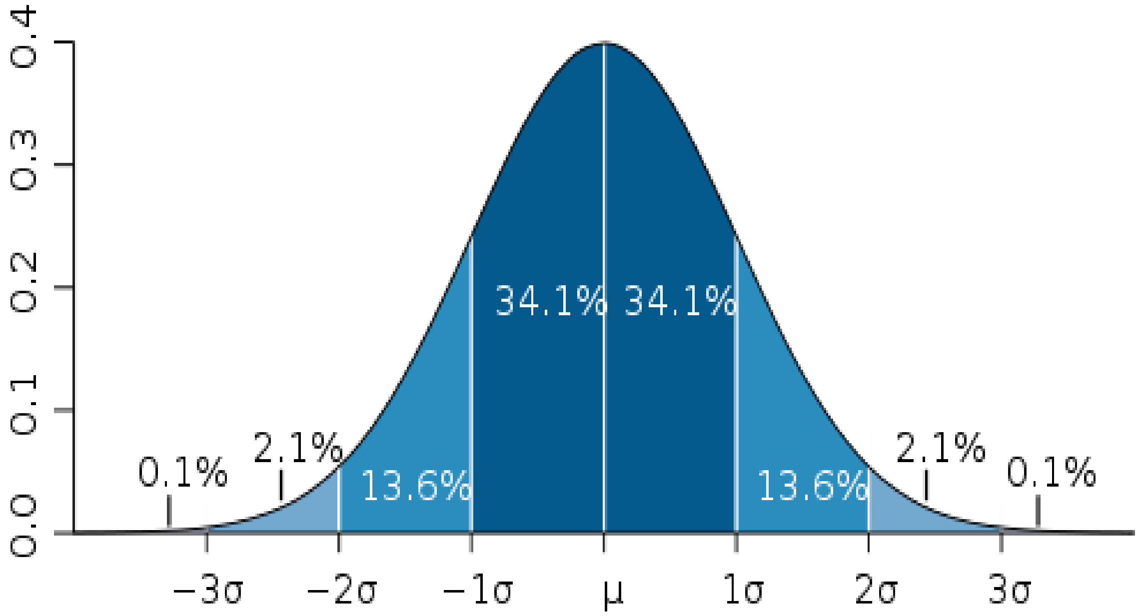
统计推断：正态分布和置信区间

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 (with a given probability) some outcomes or extrapolate relationships

统计方法主要用于预测

- 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 uses of the Gaussian distribution
  - = the 'normal' or 'bell curve'

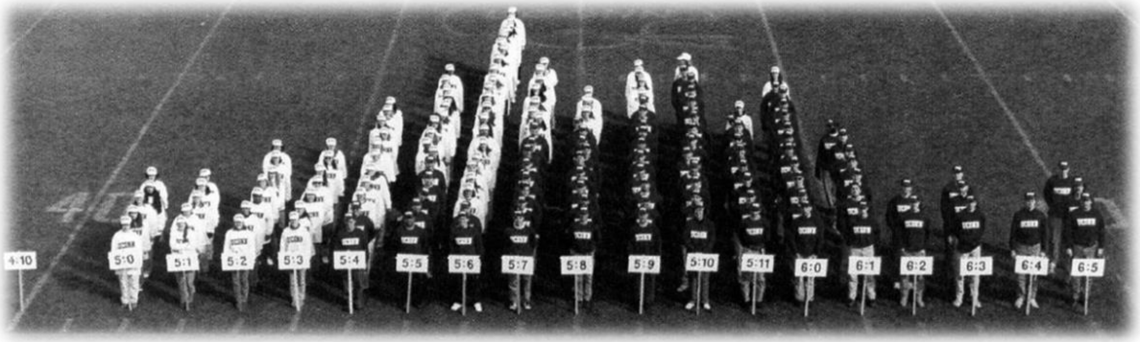
高斯分布又名正态分布
" 钟形曲线 "

Reasons for using the normal distribution

正态分布

- Many characteristics of populations have ‘bell-shaped’ distributions

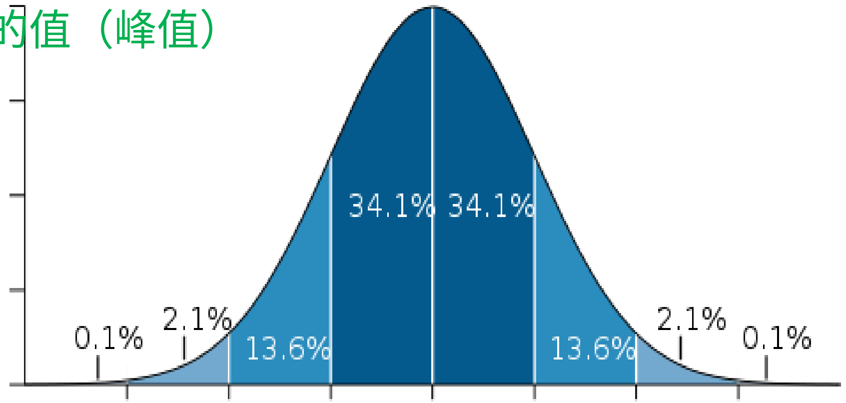
- Biological, social, economic etc. traits are often bell-shaped



The normal distribution

- The *normal distribution* is an equation that produces a bell-shaped curve; its main features are:
  - x axis shows the values of a variable, and y axis their probabilities or frequencies y轴是概率或频率
  - mean value is the most likely value (=peak) 平均值是最有可能的值（峰值）
  - probability of a value decreases with distance to mean (on either side) 一个值的概率随着距离到平均值的增加而降低（在两边）
  - sum of all probabilities is 100% (=the whole sample)

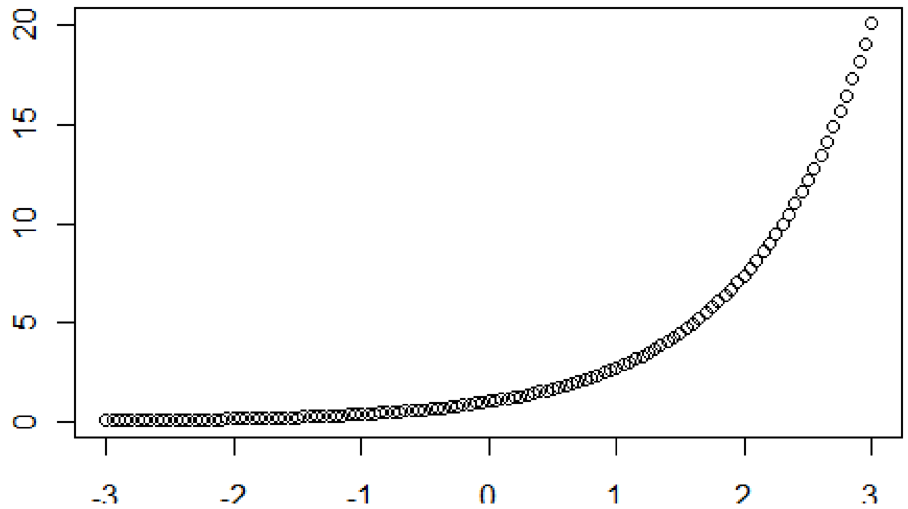
    所有概率和为100%



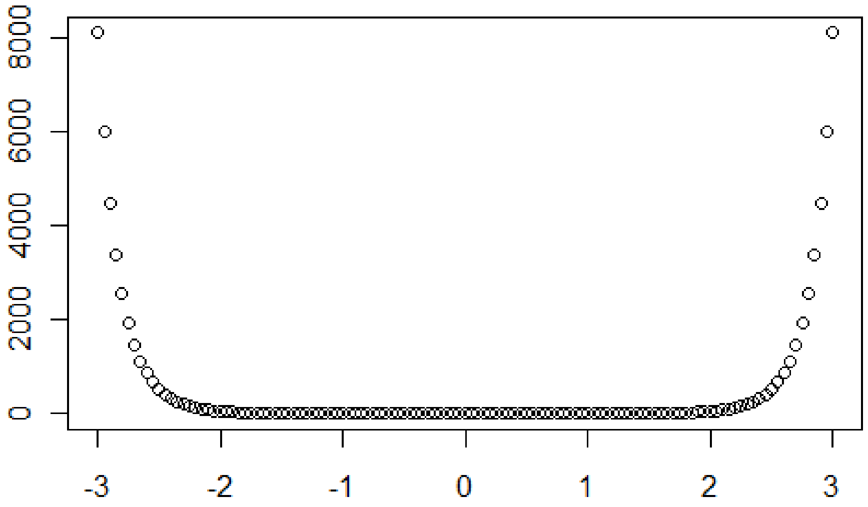
- What kind of curve/distribution produces a bell-shaped curve?
- Let’s try some exponential curves
  - =curves where $y = e^{f(x)}$

plot(exp(x) ~ x, type=“l”)

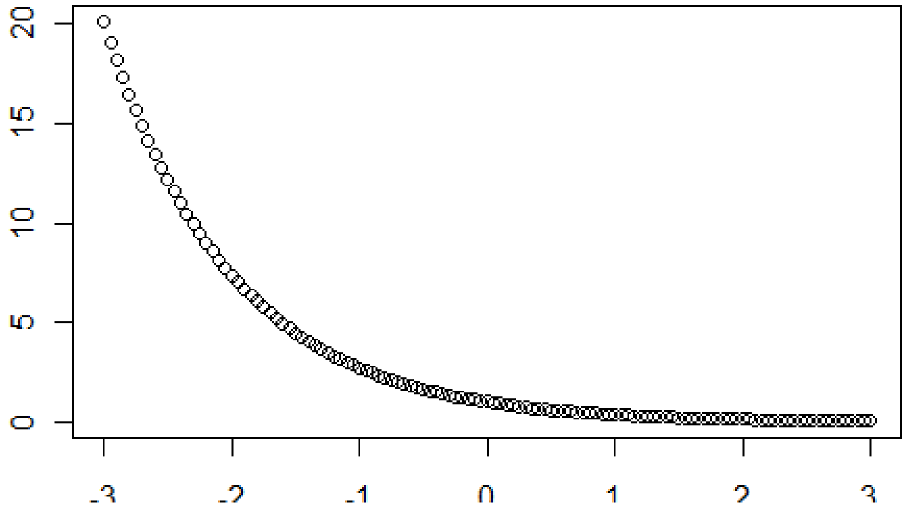
l是直线，p是点



plot(exp(x^2) ~ x, type=“l")

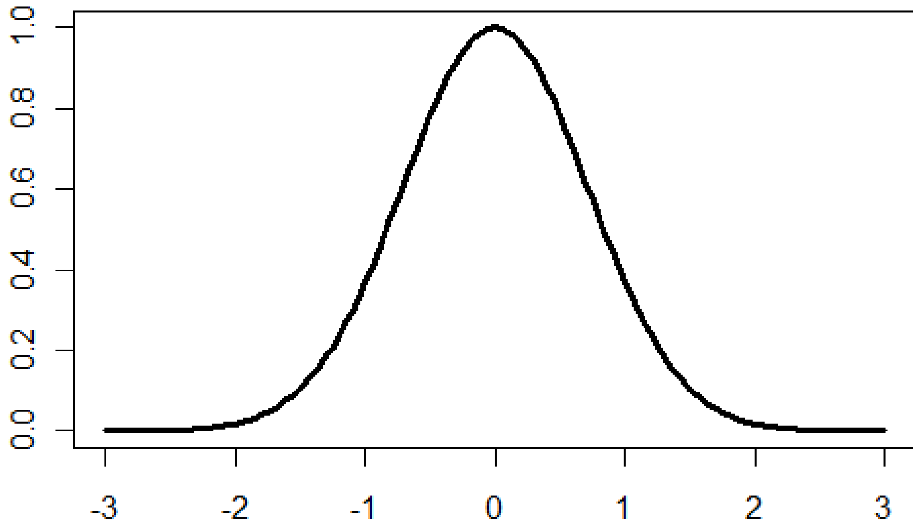


plot(exp(-x) ~ x, type=“l”)



e的-x^2次方

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

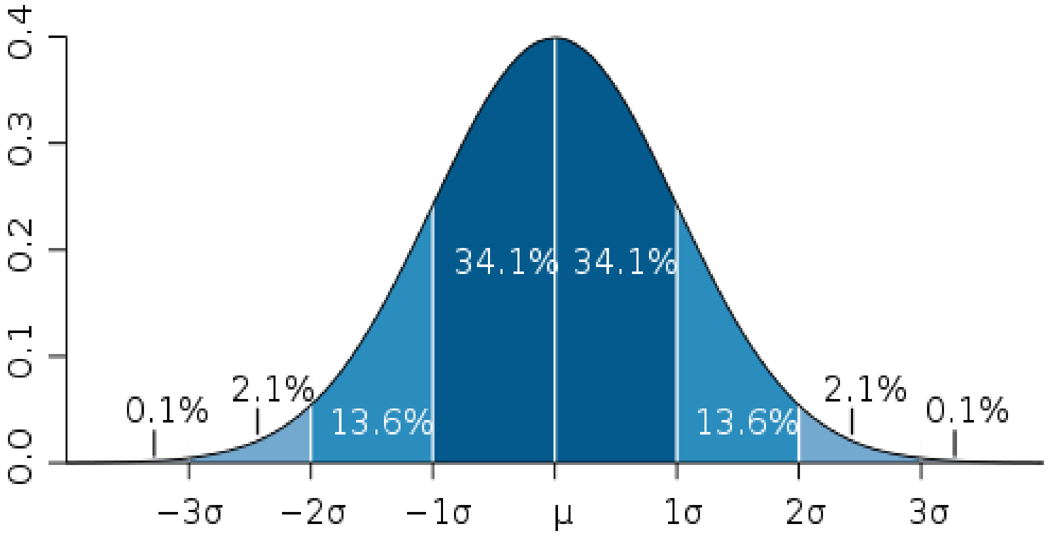


The normal equation

- The equation $y = e^{-x2}$ works and produces a bell-shaped distribution

- The normal or Gaussian curve is just a version of our curve:

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



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

- Statisticians analytically calculated probabilities and intervals from normal curve to produce tables
  - For example, we know that the probability of being over +3 sd from mean is 0.1%

Standardisation: everything is ‘normal’

真实情况下很少有均值为0和标准差为1

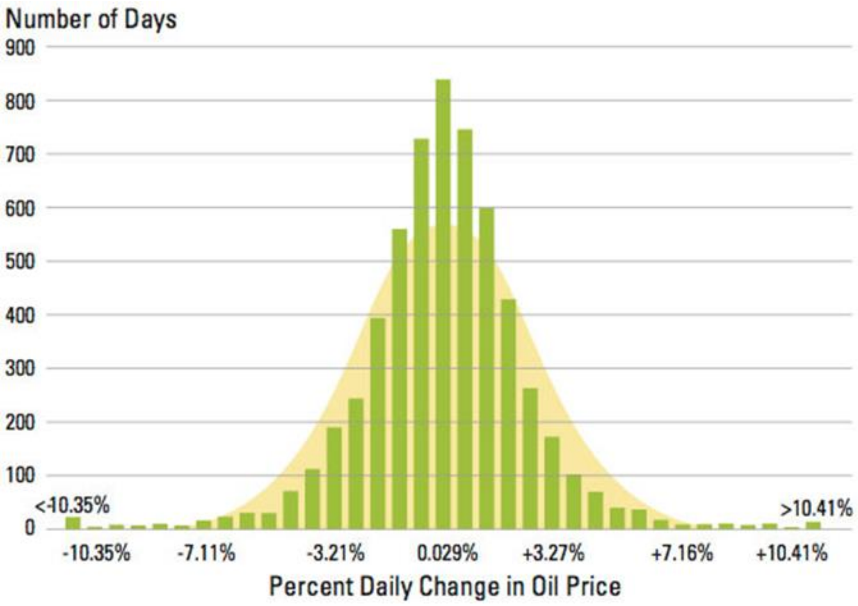
- Real traits rarely have mean=0 and standard deviation=1

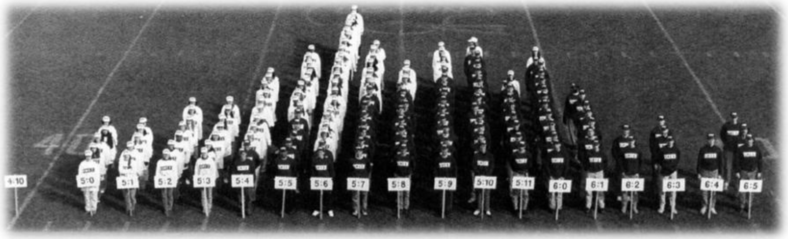
可以标准化

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

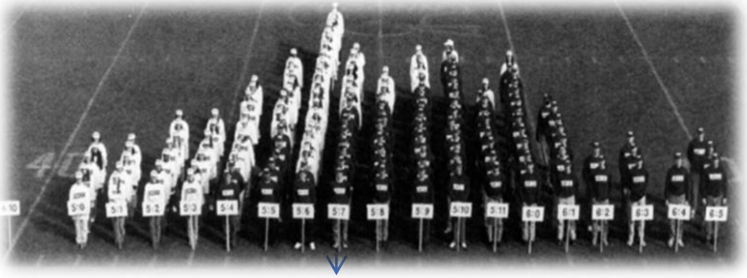
- How is this done? With ***z-scores***

又叫standard score，用于评估样本点到总体均值的距离
用于测量原始数据与数据总体均值相差多少个标准差





Calculating z-scores



Example: let's say that in a sample the mean height is mu=180cm, and sd=10cm:

For each case in your sample:

- 1) *Subtract mean* value
  - a 170cm-tall person now measures

  170-180 = -10cm (=residual)

- 2) Divide all residuals by standard deviation
  - if sd (σ, sigma) = 10 cm and mean = 180cm:
  - person measuring 170cm deviates by

  -10cm/10cm = -1 standard deviation below the mean

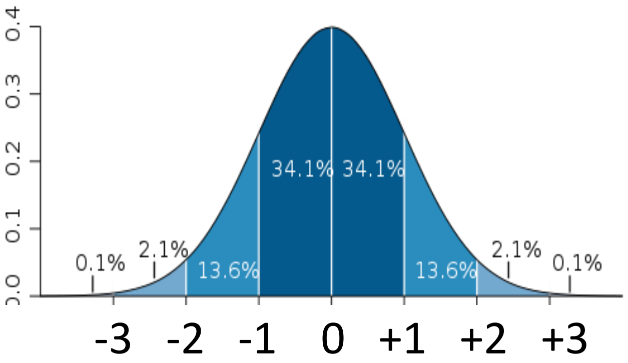
(个人值-均值)/标准差

比如小明考90分，平均成绩为95，sd=2
则z=(90-95)/2=-2.5
指小明的成绩低于班级平均分2.5个标准差

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

用z进行数据标准化，产生均值为0方差为1，无量纲的数据

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



高于平均值2.3个标准差
身高为23+180=203

Exercises:

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

(162-180)/10=-1.8

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

Intervals and cumulative probability

间隔和累计概率

- But we are more interested in *intervals* of the normal curve than individual points

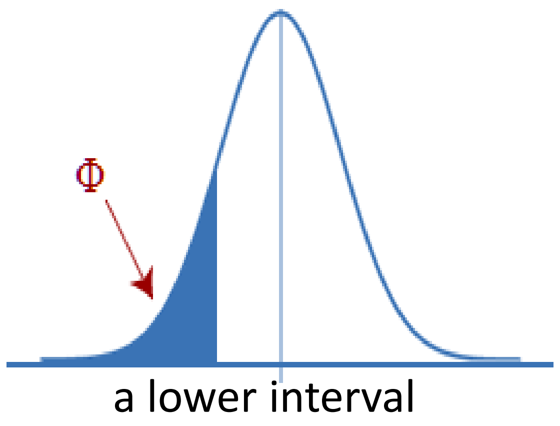
我们更感兴趣的是正态分布的间隔而不是单个的点

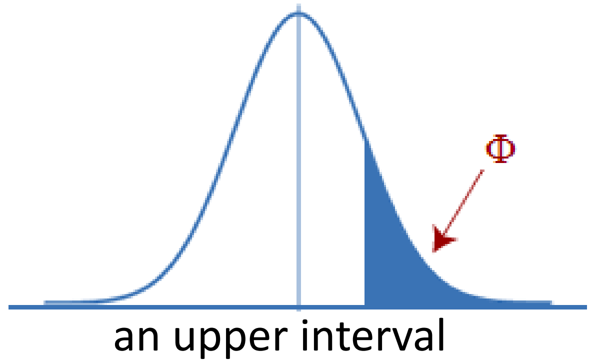
- Why? What does it mean to ask ‘what is the probability of being a millionaire in the UK?’

- It doesn’t mean the probability of having *exactly* £1 million (=a point):
    - a millionaire is someone with *£1 million or over* (=an interval)

累计概率是一个值区间的概率

- *Cumulative probability* is the probability of an interval of values





Estimating cumulative probability: lower intervals

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

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

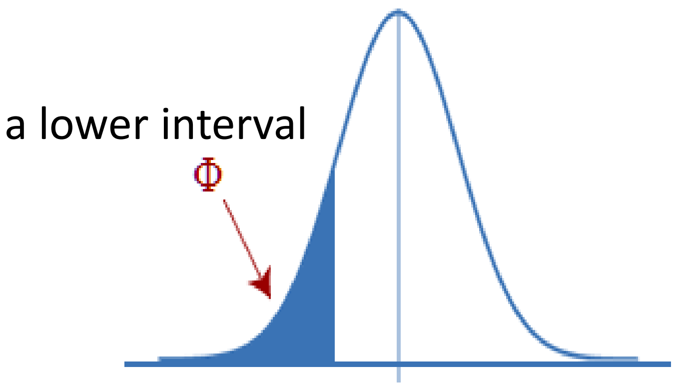
pnorm算的是从零到第一个数值的百分数(准确说是从负无穷到第一个值)，第二个数是中位数，
第三个是标准差

如果需要算一些奇怪的范围，需要动用数学知识各种加减乘除

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

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

=15.9%



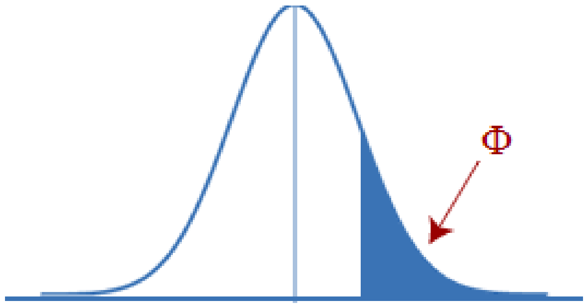
(in this case, probability of being 1 sd below mean)

Upper intervals

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

```
1-pnorm(185,180,10)
```

an upper interval



Exercise:

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

b) What is the z-score of this individual?

Probability of being ‘extreme’

极端的累积概率(这个极端大概指是两边的概率笑死了)

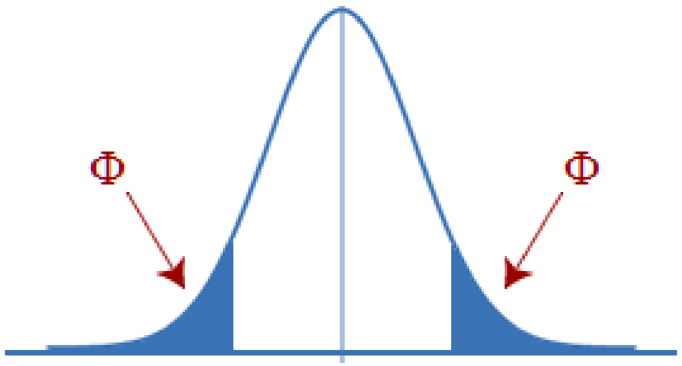
- We can also calculate probability of extreme values (i.e. too large or too small)

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

Exercise:

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

b) Which are the two z-scores? What is the interval defined by them?



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

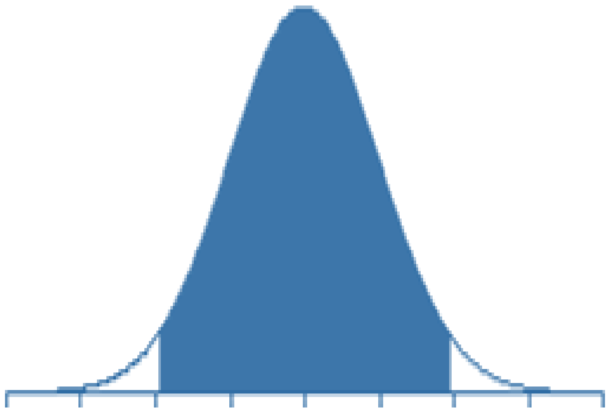
(***our most important example***)

```
pnorm(199,180,10)-pnorm(161,180,10)
```

Exercise:

a) If mean = 180cm and sd= 10cm, what is the probability of someone being between 161cm and 199cm?

b) Define the interval in terms of z-scores



Statistical testing

- In order to proceed to prediction and statistical testing, we need to define *confidence intervals*

置信区间是 “ 可接受的 ” 变化范围　包括与均值或期望值相差不大的值的区间

- Confidence intervals are ‘acceptable’ ranges of variation, i.e. intervals including the values not differing *too much* from a population mean or expected value

是基于传统定义的 “ 误差范围 ” 来确定 “ 太多 ” 意味着什么

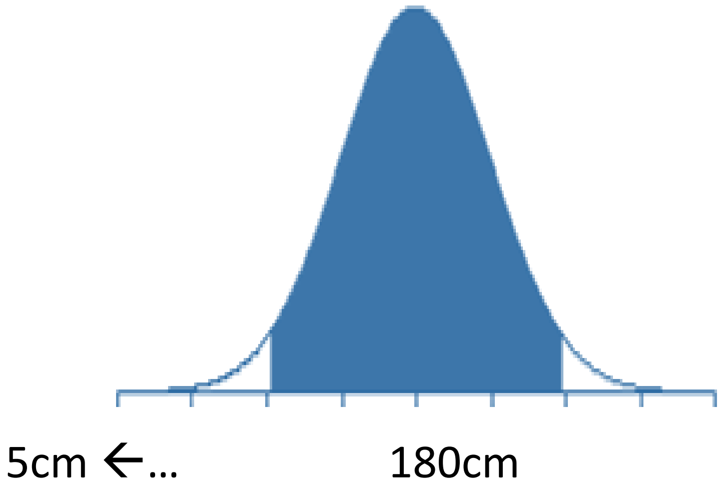
- Confidence intervals are based on conventionally-defined ‘margins of error’ establishing what ‘*too much*’ means

From ‘rare’ to ‘not one of us’

- Suppose someone tells you that they’ve found 5cm-tall people on an unknown island
  - Would you believe that??

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

- If our reference population still 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
```

如果概率很小，他们发现的生物很可能不是人类，他们不属于我们的样本和范围

- 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*

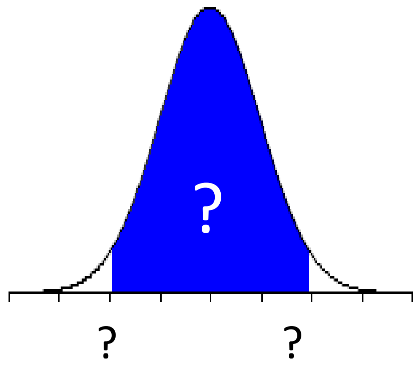
- (but bear in mind: if you are using the normal curve, a probability can be small, but never *zero*!)

用正态分布曲线，一个概率的可能性会很小，但永远不会为0

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 still 'human'

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



Question is: where, between 16% and

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

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

置信区间范围的确定是很难的（因为这个世界总有例外

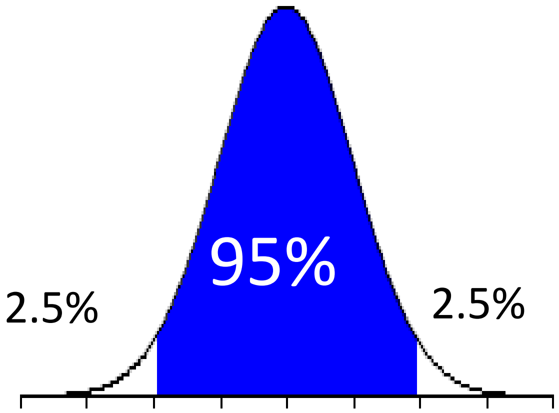
…to 95% confidence interval

- Answer: there is no objective limit 大约95%的测量值落在均值附近的±2倍sd之间
  - limit is set *conventionally* 界限是公认(惯例)设定的

- Most often, boundary is set at 5% 通常情况下是5%
  - or less frequently, 1%
  - then, if a probability is > 5%, i.e. within a 95% confidence interval around mean, it is accepted as part of that distribution; it is not 'rare' (not too small, not too large)
  - if probability of a value is < 5%, it is too 'rare'; it is defined as not in the distribution

通常情况下是95%的置信区间

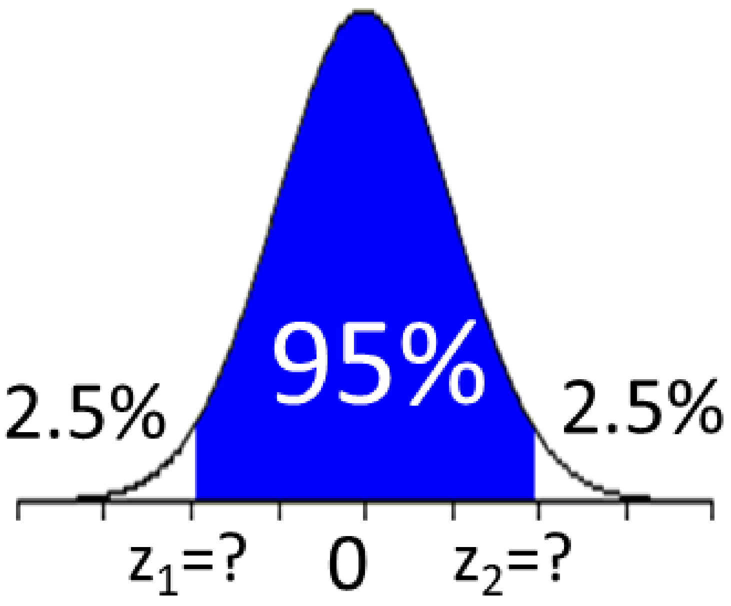
- The conventional value of 5% defines a 95% confidence interval
  - it excludes 2.5% cases on each side, i.e. too low or too high, as not belonging in the distribution
  - It defines confidence or belief that the case belongs in the distribution



Boundaries of the 95% CI

qnorm(0.025)是-1.959964

- If we define our CI at 95%, how much do you need to deviate from the mean to be in the ‘extremely rare’ 5%?



Exercise:

Estimate approximate lower boundary z_1 and upper boundary z_2 of the standardised 95% CI

Using the *pnorm* function and trial and error!

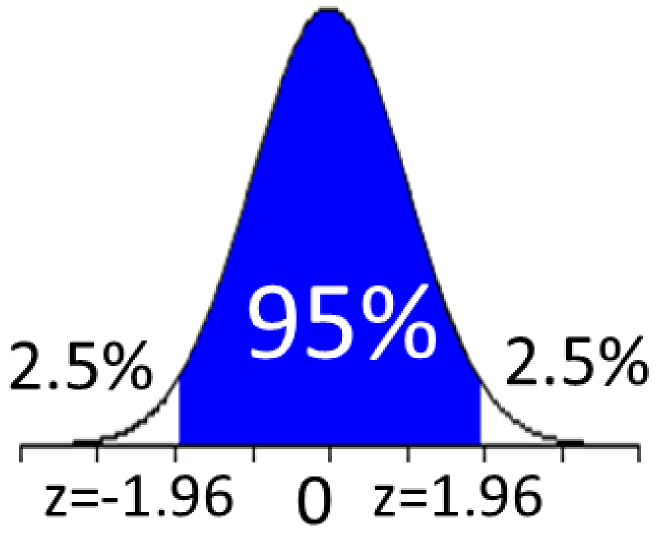
Present the values in

- z-scores
- cm (assuming mean = 180cm and sd = 10cm)

Boundaries of the 95% CI

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

±1,96



Note:

- This is true for large sample sizes
- Values change as a function of sample size
  - if samples are small, z-scores defining lower and upper boundaries are smaller

如果样本很小，定义下限和上限的z分数会更小

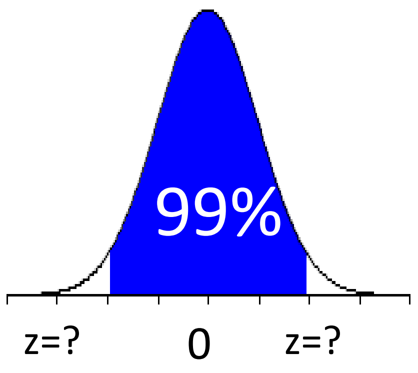
d 概率密度函数
p 分布函数
q 分布函数的反函数
r 产生相同分布的随机数
正态分布的英文是normal distribution，所以与正态分布相关的函数都是以norm结尾的

rnorm(n=100,mean=15,sd=2)随机生成均值为15方差为2满足正态分布的100个数，使用round函数取整

Exercise:

Estimate approximate lower and upper boundaries of the **99% CI** qnorm(0.005)

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

```
#Coursework
#import KungCensus
#creating a file with adult women only:
kungadultfemales <- subset(KungCensus, age > 18 & sex == 'woman')
#or if you don't want to create a new file but only filter cases for the command
hist(KungCensus$weight[KungCensus$age > 18  & KungCensus$sex == 'woman'])
```

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

Use new file or:

> hist(KungCensus$weight[KungCensus$age > 18 & KungCensus$sex == “woman”])

c) How many adult females with missing weight data?

```
length(kungadultfemales$weight)
```

Tip: function *summary*

d) How many adult females with weight data?

```
> mean(kungadultfemales$weight,na.rm = T)
[1] 42.07432
> sd(kungadultfemales$weight,na.rm = T)
[1] 5.301835
```

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

```
> pnorm(40,mean(kungadultfemales$weight,na.rm = T),sd(kungadultfemales$weight,na.rm = T))
[1] 0.3478077
```

ii) over 60 kg

```
> 1-pnorm(60,mean(kungadultfemales$weight,na.rm = T),sd(kungadultfemales$weight,na.rm = T))
[1] 0.0003610699
```

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

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

```
> pnorm(-3,0,1)或者直接pnorm(-3)
[1] 0.001349898
> 1-pnorm(3,0,1)
[1] 0.001349898
```

- Some answers to final exercises:

1c) 68

1d) 264 – 68

2a) 0.001349898

2b) 0.001349898

2c) 99.73% CI