17C Laboratory & Professional Skills: Data Analysis

Emma Rand
Data Analysis in R

Week 4: Normal distributions, calculating probabilities and Confidence Intervals

Last week

- types of variable: Discrete (categories and counts) or Continuous
- the logic of hypothesis testing
- In RStudio:
  - reading in data files
  - Working directories and paths
  - summarising and plotting data.
  - saving figures and laying out a report in Word.

Summary of this week

- Continuous variables: The normal distribution. Because it is the basis of many tests (parametric tests such as *t*-test, regression and ANOVA)
  - Properties of normal distributions
  - Sampling distribution of the mean and the standard error
  - Confidence intervals
- In RStudio
  - Calculate probabilities and quantiles from normal distributions
  - Calculate confidence intervals

Learning objectives for the week

By actively following the material and carrying out the independent study the successful student will be able to:

- Explain the properties of ‘normal distributions’ (MLO 1 and 2)
- Define the sampling distribution of the mean and the standard error (MLO 1 and 4)
- Explain what a confidence interval is (MLO 1 and 4)
- Calculate probabilities and quantiles and in R (MLO 3 and 4)
- Calculate confidence intervals for large and small samples in R (MLO 3 and 4)

To understand confidence intervals

We need to understand the standard error

To understand the standard error we need to understand the sample distribution of the mean

To understand the sampling distribution of the mean we need to understand a distribution

To understand confidence intervals

We need to understand the standard error

To understand the standard error we need to understand the sample distribution of the mean

To understand the sampling distribution of the mean we need to understand a distribution

1. We will learn what the normal distribution is

To understand confidence intervals

We need to understand the standard error

To understand the standard error we need to understand the sample distribution of the mean

2. We will learn what the sample distribution of the mean is

To understand the sampling distribution of the mean we need to understand a distribution

1. We will learn what the normal distribution is

To understand confidence intervals

We need to understand the standard error

3. We will learn what the standard error is

To understand the standard error we need to understand the sample distribution of the mean

2. We will learn what the sample distribution of the mean is

To understand the sampling distribution of the mean we need to understand a distribution

1. We will learn what the normal distribution is

What is the normal distribution?

What do we mean by distribution?

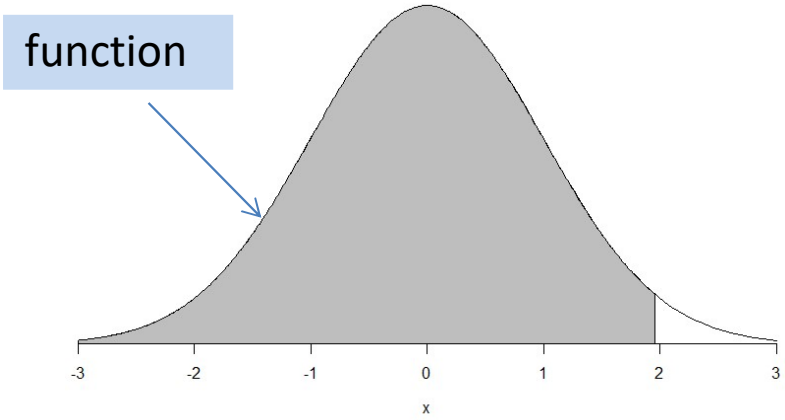
What do we mean by distribution?

- any variable has a distribution
- A distribution is the values a variable can take and the chance of them occurring
- Distribution is a function (relationship)
- Parameters tune the shape of the distribution

What do we mean by distribution?

What do we mean by distribution?

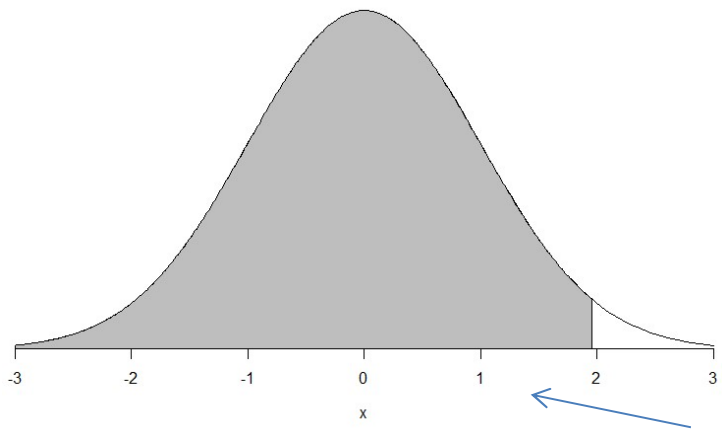
- any variable has a distribution
- A distribution is the values a variable can take and the chance of them occurring
- Distribution is a function (relationship)
- Parameters tune the shape of the distribution



What do we mean by distribution?

What do we mean by distribution?

- any variable has a distribution
- A distribution is the values a variable can take and the chance of them occurring
- Distribution is a function (relationship)
- Parameters tune the shape of the distribution

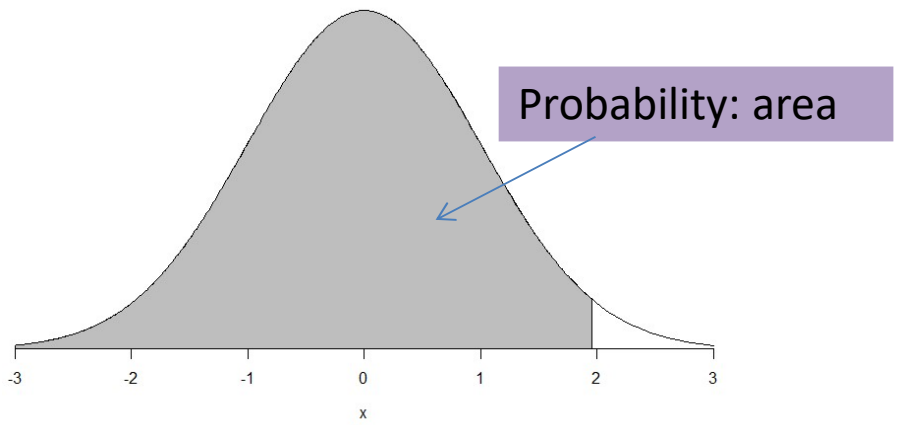


Value a variable can take

What do we mean by distribution?

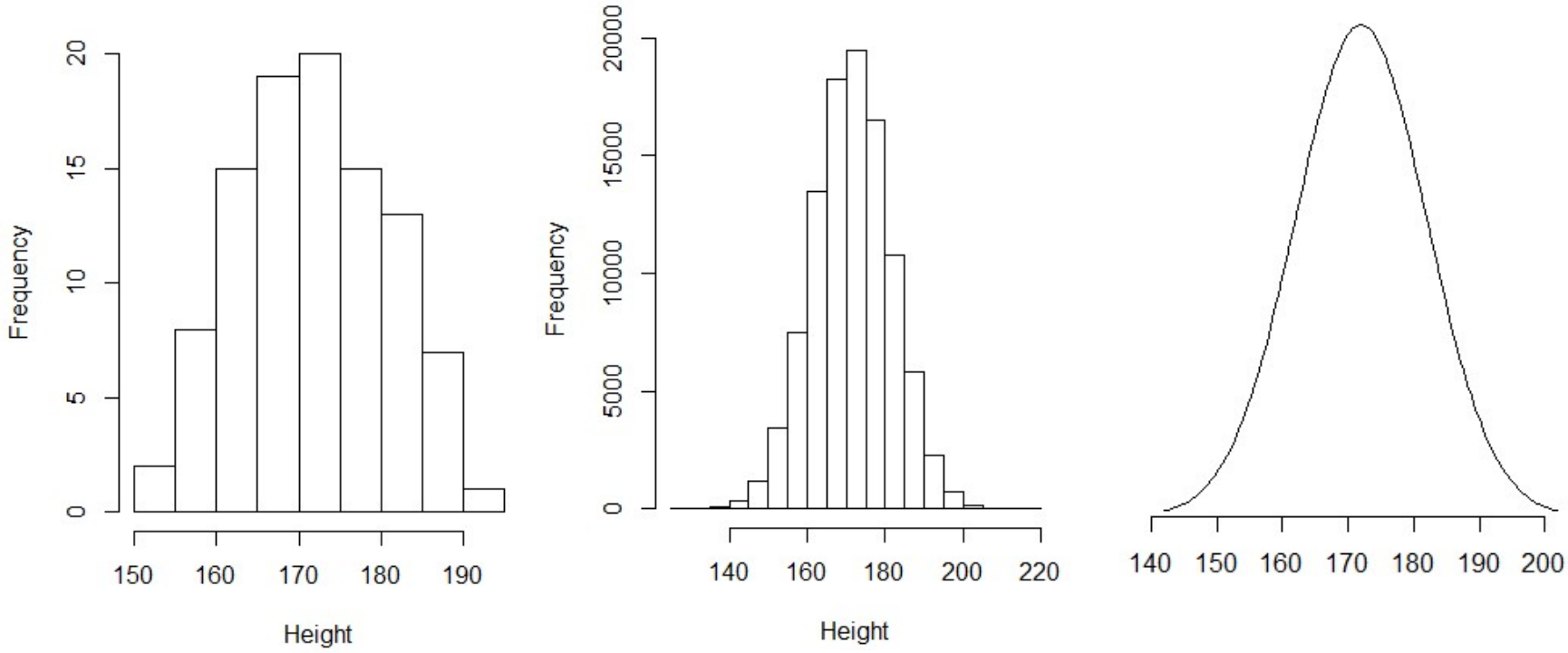
What do we mean by distribution?

- any variable has a distribution
- A distribution is the values a variable can take and the chance of them occurring
- Distribution is a function (relationship)
- Parameters tune the shape of the distribution



The normal distribution

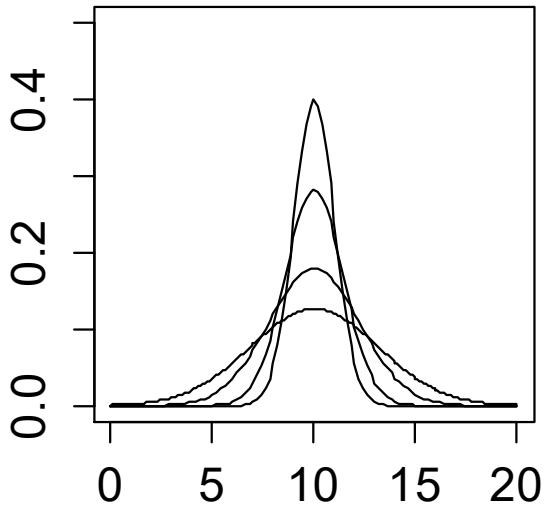
E.g., height, length, concentration



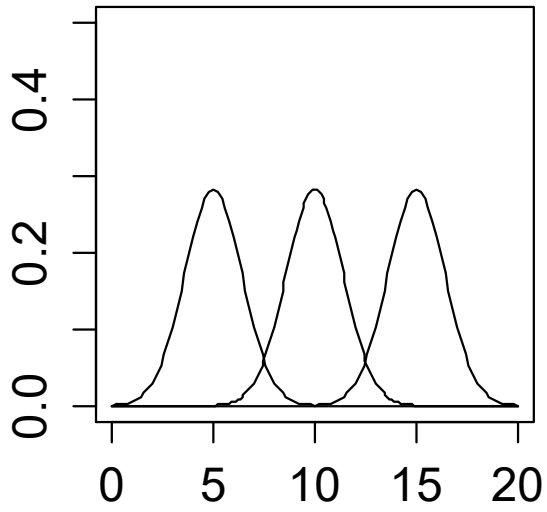
The normal distribution

Can vary in two ways – 2 parameters

Variance – how wide?

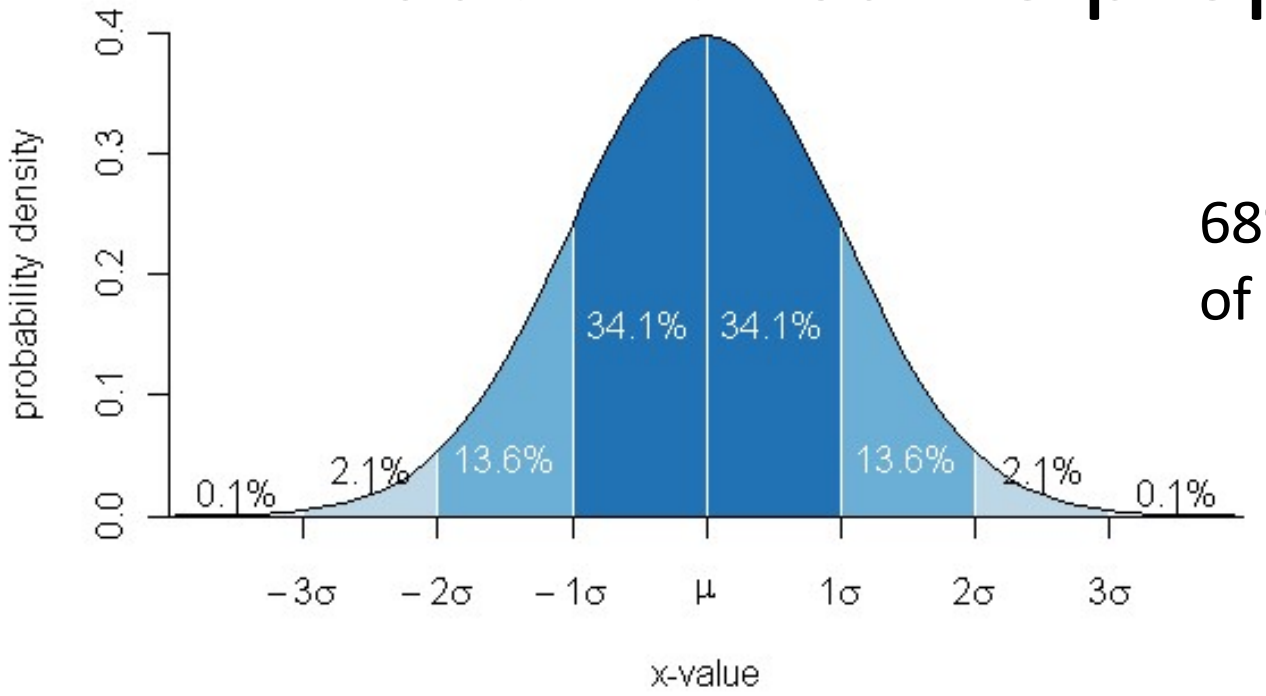


Mean – where on the axis?

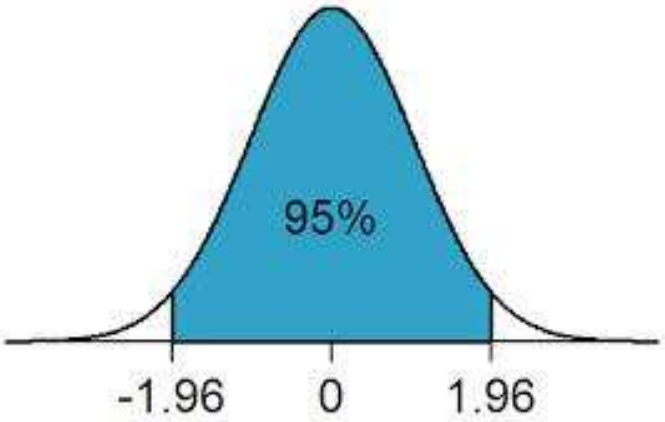


The normal distribution

….but with same properties



68% within 1 SD of the mean

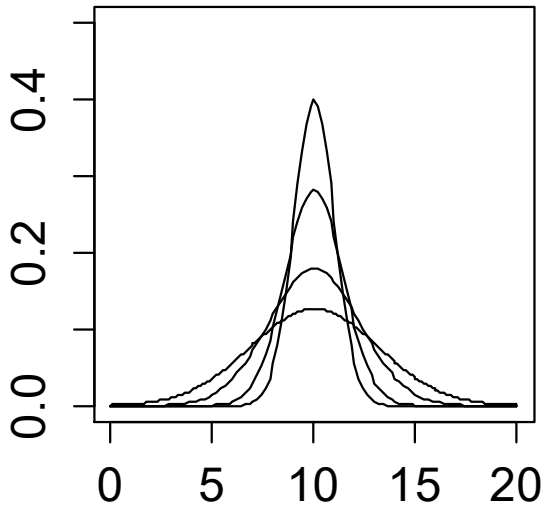


95% of observations are within 1.96 sd of the mean

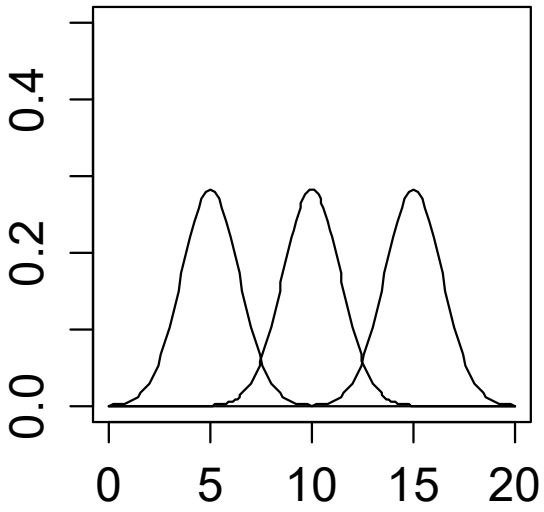
The normal distribution

Can vary in two ways – 2 parameters

Variance – how wide?



Mean – where on the axis?



The normal distribution

The mean

- Population mean

  μ (mu) in <u>whole</u> population

  There is a true value for the mean if you measured every individual

- Sample mean

  $\bar{x}$ (x bar) in <u>sample</u>

  You don't measure every individual, you measure some (a sample). $\bar{x}$ is an estimate of μ

$$\bar{x} = \frac{\sum_i x_i}{n}$$

The normal distribution

The variance

In a sample, each sample value differs from the mean.
Each difference is called a deviation (or a residual)
"average of the squared deviations from the mean"

- Sample variance

  s^2 (s-squared) in <u>sample</u>

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

You need to understand the concept rather than remember the formula

The normal distribution

The standard deviation

"The average of the (absolute) deviations from the mean"

- the square root of the variance
- Sample standard deviation: *s*
- Tells you how variable the <u>values</u> are

95% of observations are within 1.96 standard deviation of the mean

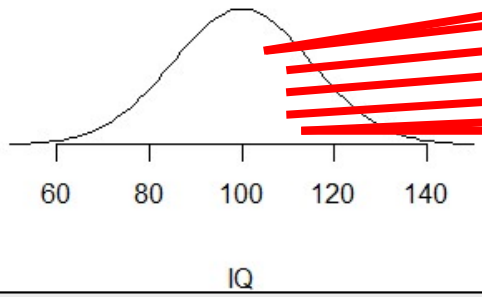
Sampling distribution of the mean and the standard error

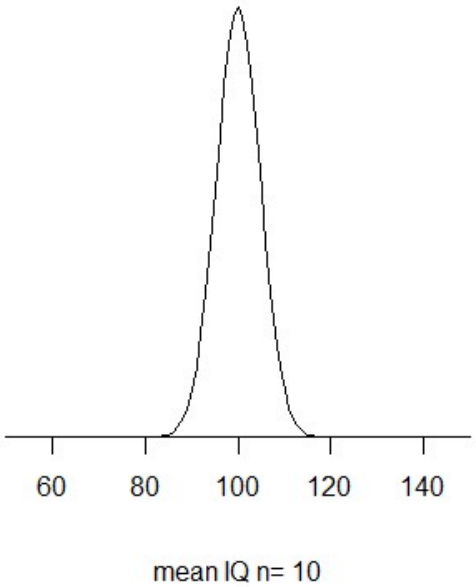
Sampling distribution of the mean

- A population has one true mean, μ
- A sample taken from that population has a mean, $\bar{x}$ that will differ from μ
- And from other sample $\bar{x}$
- That is, $\bar{x}$ has a distribution

Sampling distribution of the mean

Many samples
All same size
e.g., n=10



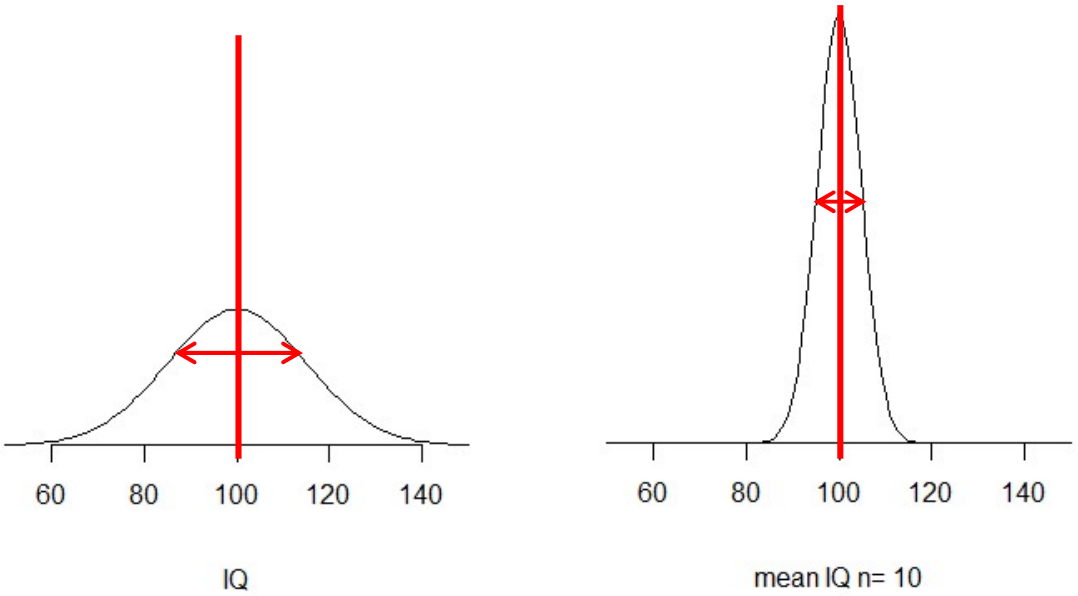


Distribution – whole population
mean = 100
sd = 15

Each sample mean differs from 100 by chance

<u>Distribution of the means</u>

Sampling distribution of the mean



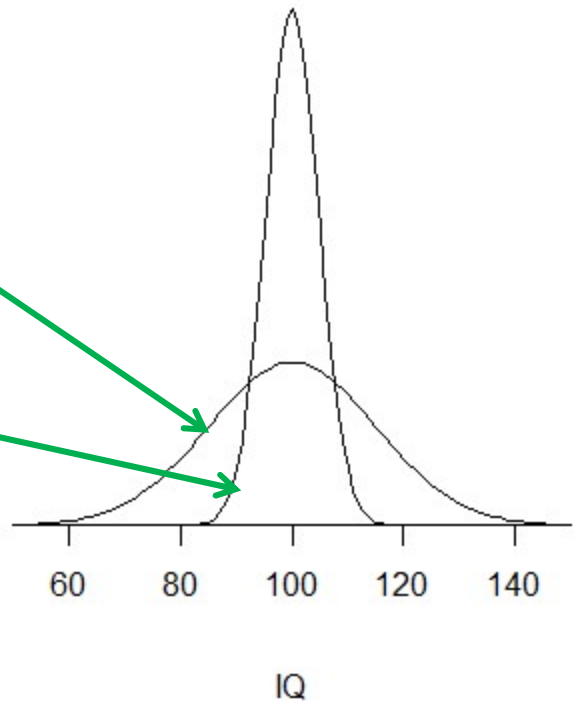
- Has the same mean as the parent
- But a different (lower) standard deviation
- And we call it the ‘standard error’

Sampling distribution of the mean

Whole population
mean = 100
sd = 15

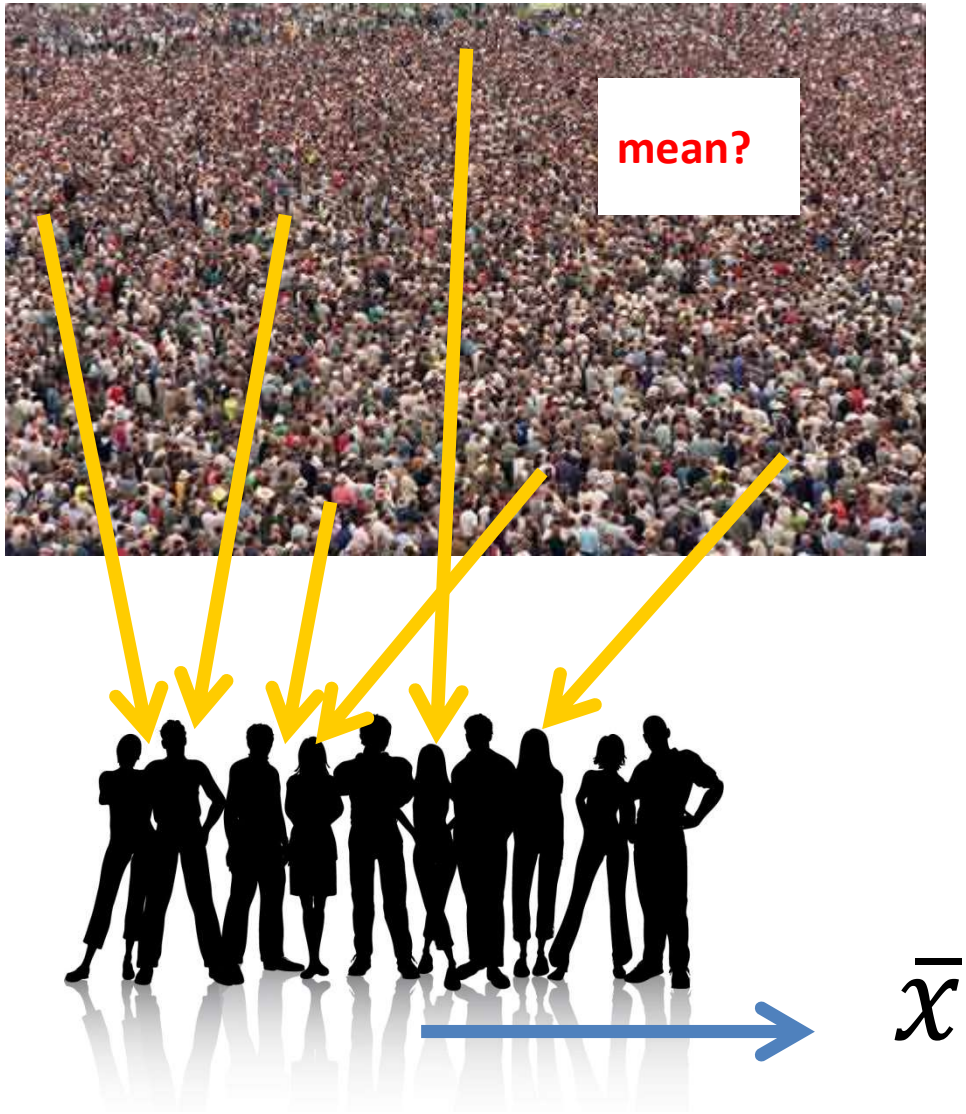
<u>Distribution of the means</u>
mean = 100
se is related to sd in predictable way

se = sd/√n



Standard error: the standard deviation of the sample means.
Tells you how variable the <u>sample means</u> are

Sampling distribution of the mean



Why does it matter?

- We usually only have samples!
- We do not know population parameters
- We use samples to estimate

……..

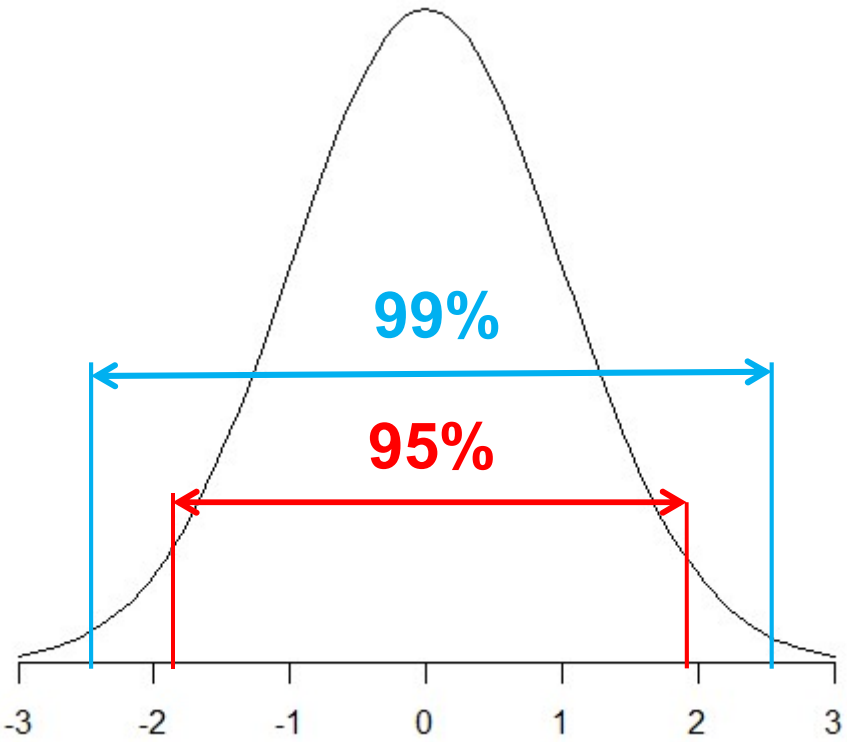
Confidence intervals

Confidence intervals

- How confident can we be that our sample mean is a good estimate of the true value?
- Confidence intervals give the highest and lowest *likely* values
- Likely means 95%, 99%, 99.9%

The normal distribution

Confidence intervals: large samples



95% of sample means are within 1.96 s.e. of the population mean

99% of sample means are within 2.58 s.e. of the population mean

Confidence intervals: <u>large samples</u>

The mean plus or minus a bit

- $\bar{x} \pm 1.96 \times s.e.$
- i.e., 95% certain population mean is between $\bar{x} - 1.96 \times s.e.$ and $\bar{x} + 1.96 \times s.e.$

Do I have to remember 1.96? Not if you have R

```
> qnorm(0.975)
[1] 1.959964
```

what is qnorm()???

Confidence intervals: large samples

pnorm() and qnorm()

The Normal Distribution

Description

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to `mean` and standard deviation equal to `sd`.

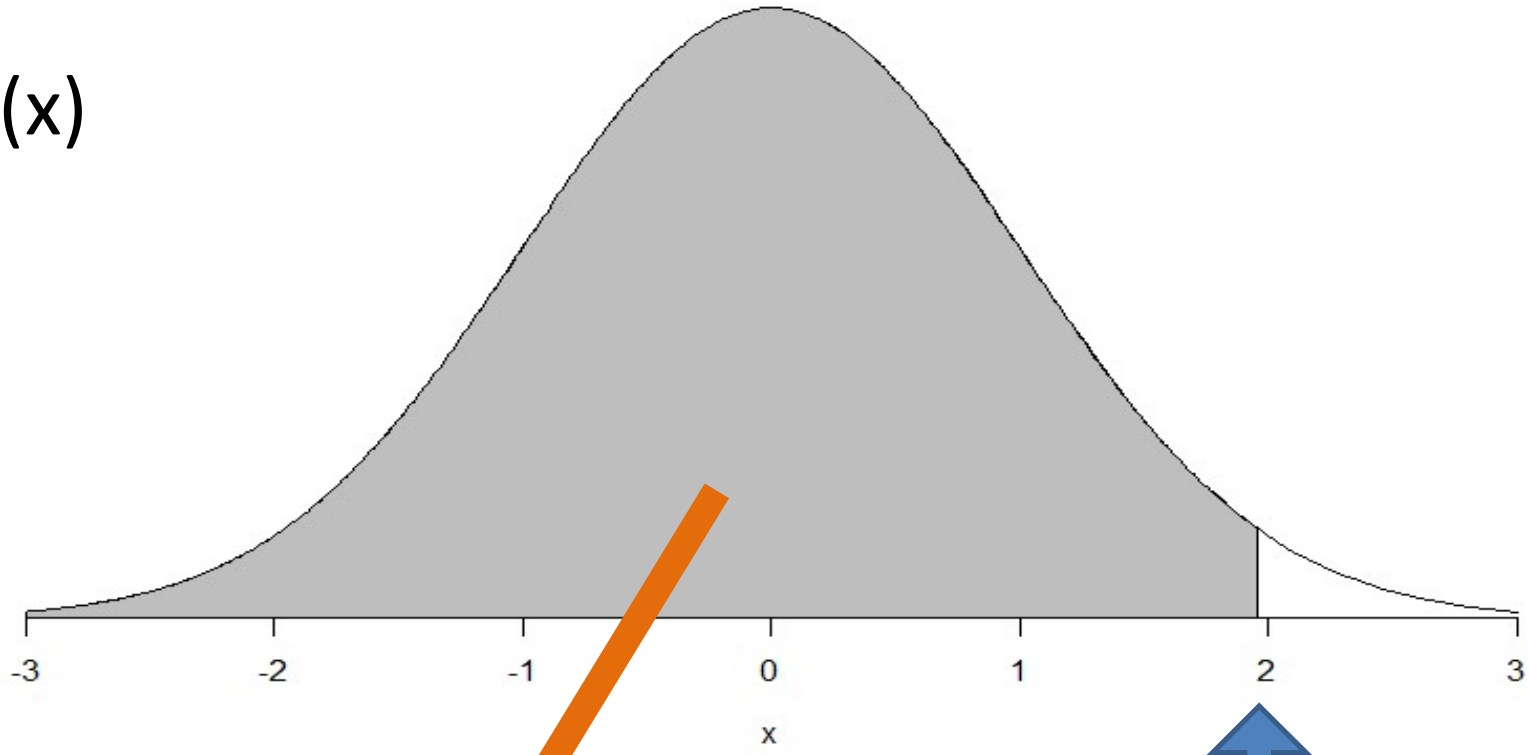
Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Arguments

Confidence intervals: large samples

pnorm(x)

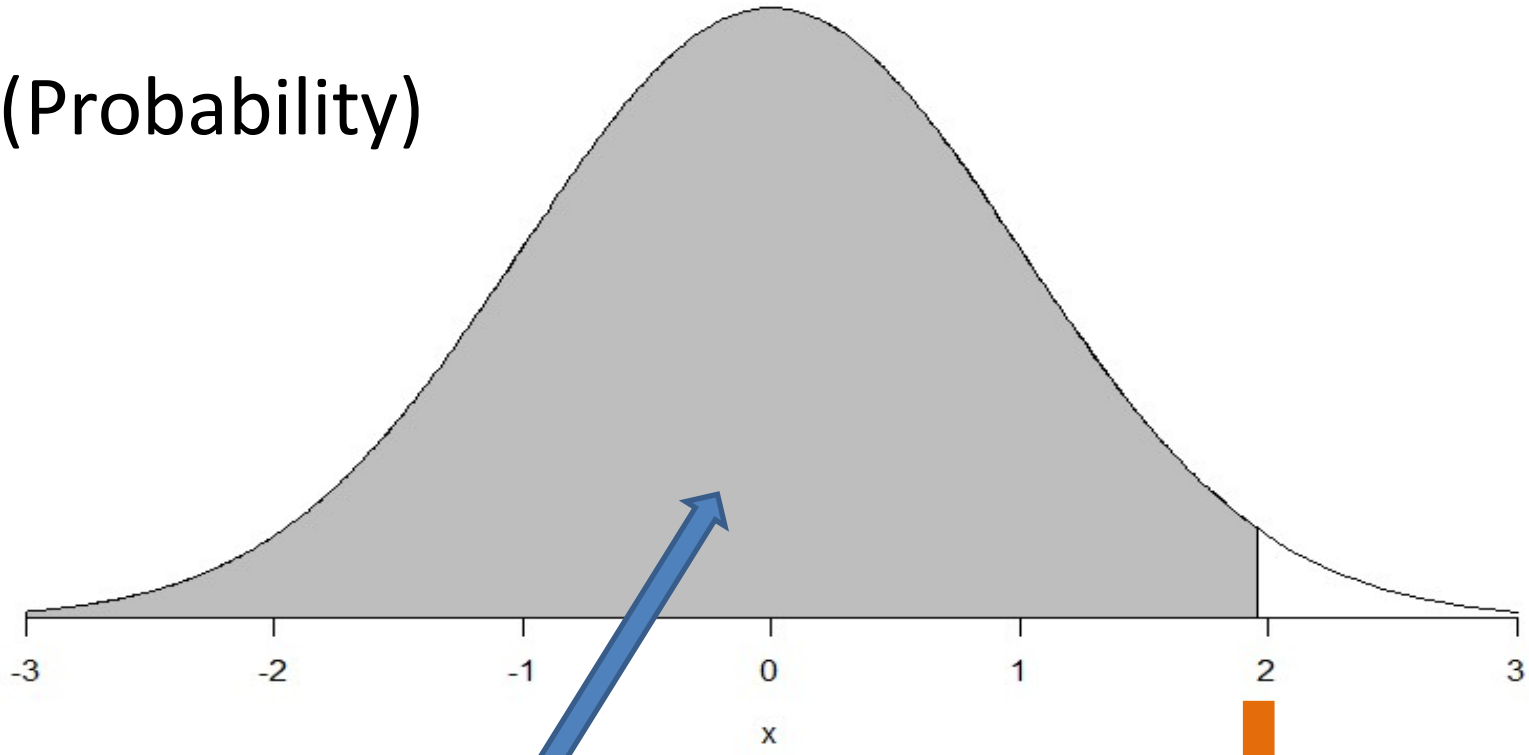


Returns probability

x

Confidence intervals: large samples

qnorm(Probability)



probability

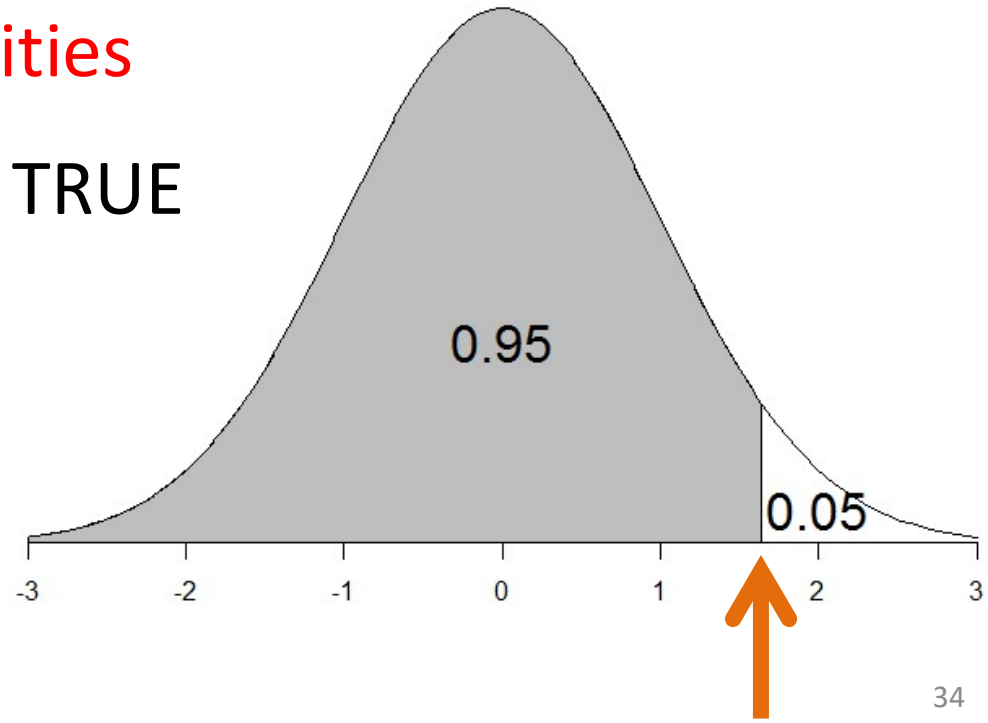
Returns x

Confidence intervals: large samples

Why qnorm(0.975) and not qnorm(0.95)?

- Because qnorm() gives 'one-tailed' probabilities
- uses the lower tail = TRUE

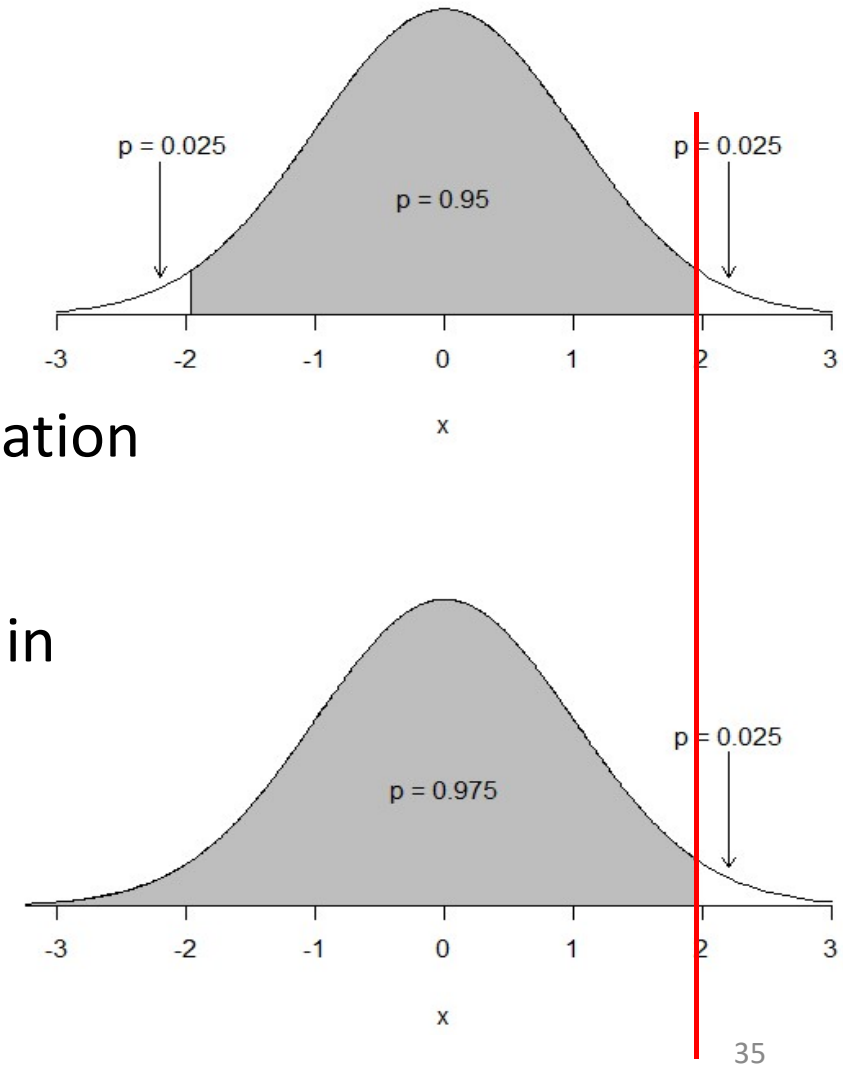
(By default)



The normal distribution

Confidence intervals: <u>large samples</u>

Why qnorm(0.975) and not qnorm(0.95)?



- We want to be 95% certain population mean is between

  $\bar{x} - 1.96 \times s.e.$ and $\bar{x} + 1.96 \times s.e.$

- We want 0.05 in both tails, 0.025 in each tail

- So we need to give it

  1- 0.025

  =0.975

Confidence intervals: large samples

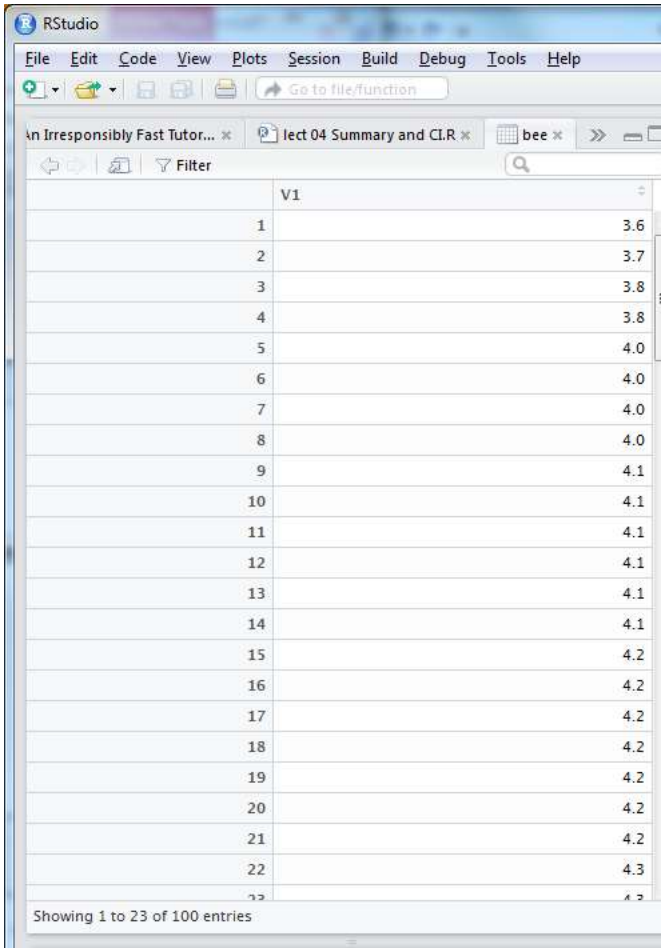
New population of honey bees – how big are their left wings?

- Take a ‘representative’ sample

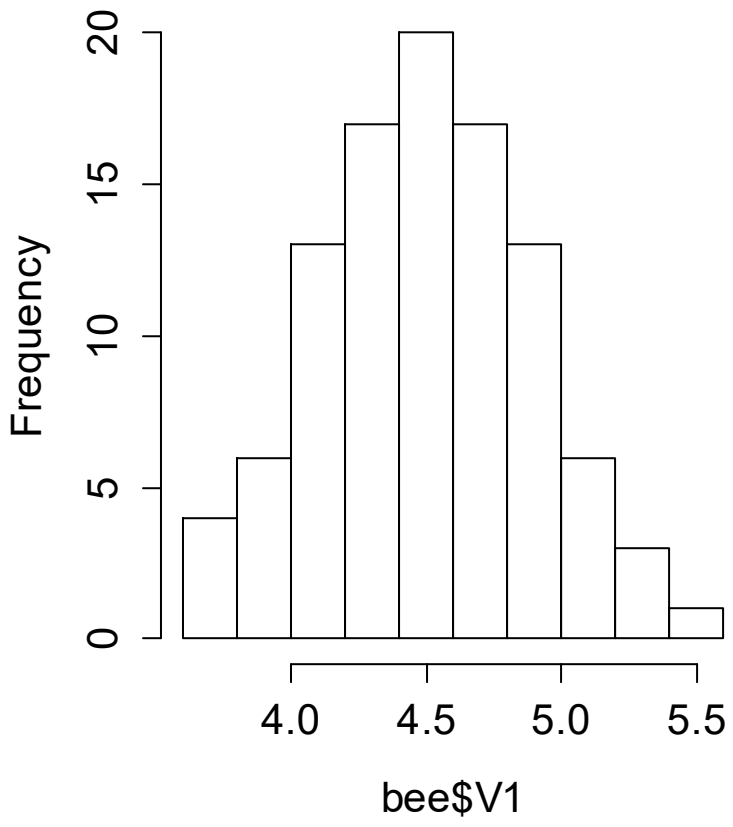
The normal distribution

Confidence intervals: large samples

Left wing widths of 100 honey bees (mm)



	V1
1	3.6
2	3.7
3	3.8
4	3.8
5	4.0
6	4.0
7	4.0
8	4.0
9	4.1
10	4.1
11	4.1
12	4.1
13	4.1
14	4.1
15	4.2
16	4.2
17	4.2
18	4.2
19	4.2
20	4.2
21	4.2
22	4.3



Confidence intervals: large samples

Left wing widths of 100 honey bees (mm) $\bar{x} \pm 1.96 \times s.e.$

Mean

```
m <- mean(bee$v1)
[1] 4.55
```

Standard error **se = sd/√n**

```
se <- sd(bee$v1)/sqrt(length(bee$v1))
[1] 0.03919647
```

quantile

```
q <- qnorm(0.975)
[1] 1.959964
```

amount to add/subtract

```
q * se %>% round(2)
[1] 0.08
```

Upper confidence limit

```
m + q * se %>% round(2)
[1] 4.63
```

Lower confidence limit

```
m - q * se %>% round(2)
[1] 4.47
```

Confidence intervals: large samples

Left wing widths of 100 honey bees (mm)

- Sample mean = 4.55 mm
- 95% certain population mean is between: 4.63 mm and 4.47 mm
- We would normally summarise as:

The 95% confidence interval on the mean was 4.55 ± 0.08 mm.

Confidence intervals: <u>small samples</u>

Use $t_{[d.f.]}$ (`qt()`) rather than 1.96 (`qnorm()`)

$$\bar{x} \pm t_{[d.f.]} \times s.e.$$

(Sampling distribution of the mean for small samples is not quite normal but instead follows a *t* distribution)

Confidence intervals: <u>small samples</u>

- Value depends on degrees of freedom
- $t_{[\infty]}$ = 1.96
- 95% of sample means $\bar{x} \pm t_{[d.f.]} \times s.e.$
- Need qt() rather than qnorm()

```
> qt(0.975, df = 4)
[1] 2.776445
> qt(0.975, df = 9)
[1] 2.262157
```

```
> qt(0.975, df = 99)
[1] 1.984217
> qt(0.975, df = 999)
[1] 1.962341
```

The normal distribution

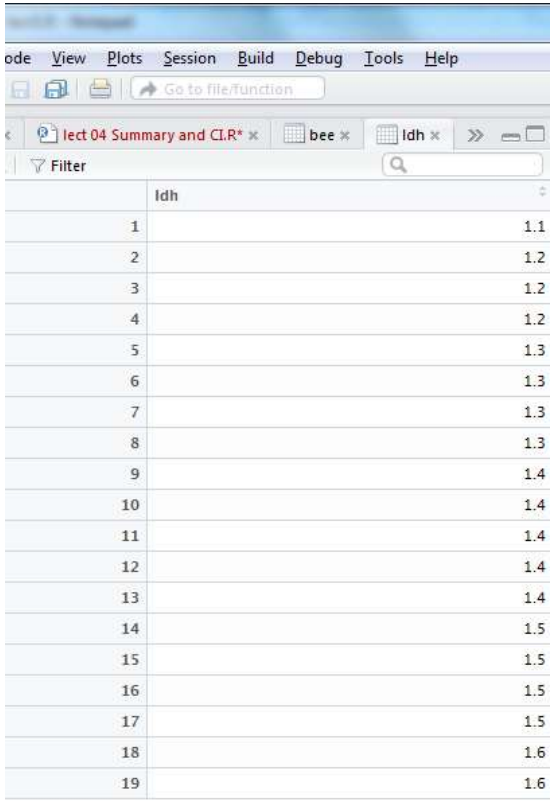
Confidence intervals: <u>small samples</u>

19 lactate dehydrogenase solutions to a recipe that <u>should</u> yield a concentration of 1.5 µmols l^{-1}

How good is the recipe/ability to follow the recipe?

Confidence intervals: small samples

19 lactate dehydrogenase solutions to a recipe that should yield a concentration of 1.5 µmols l^{-1}



	ldh
1	1.1
2	1.2
3	1.2
4	1.2
5	1.3
6	1.3
7	1.3
8	1.3
9	1.4
10	1.4
11	1.4
12	1.4
13	1.4
14	1.5
15	1.5
16	1.5
17	1.5
18	1.6
19	1.6

```
mean(ldh$ldh)
[1] 1.373684
```

Mean of sample is 1.37 µmols l^{-1}

What is an estimate the population mean?

The normal distribution

Confidence intervals: <u>small samples</u>

$$\bar{x} \pm t_{[d.f.]} \times s.e.$$

Mean

```
m <- mean(ldh$ldh);m
1.373684
```

Standard error

```
se <- sd(ldh$ldh)/sqrt(length(ldh$ldh)); se
[1] 0.03230167
```

qt – need df

```
df <- length(ldh$ldh) -1 ; df
[1] 18
t <- qt(0.975, df = df); t
[1] 2.100922
```

Upper CL

```
round(m + t * se, 2)
[1] 1.44
```

Lower CL

```
round(m - t * se, 2)
[1] 1.31
```

Confidence intervals: small samples

19 lactate dehydrogenase solutions to a recipe that should yield a concentration of 1.5 µmols l^{-1}

The 95% confidence interval on the mean was 1.37 ± 0.07 µmols l^{-1}.

- 95% certain population mean is between: 1.31 and 1.44 µmols l^{-1}
- What does this tell us about the recipe/ability to follow recipe?

Summary

- Normal distributions are common
- They have two parameters: the mean and standard deviation
- All normal distributions have the same properties so we can use them for probabilities and CI
- The standard error is the standard deviation of the sample means
- pnorm and qnorm are each others inverse; give the probability and the quantile respectively; have lower.tail = TRUE by default
- CI for large samples: $\bar{x} \pm 1.96 \times s.e.$
- CI for small samples: $\bar{x} \pm t_{[d.f.]} \times s.e.$