Statistical Inference

Bob O'Hara

bob.ohara@ntnu.no

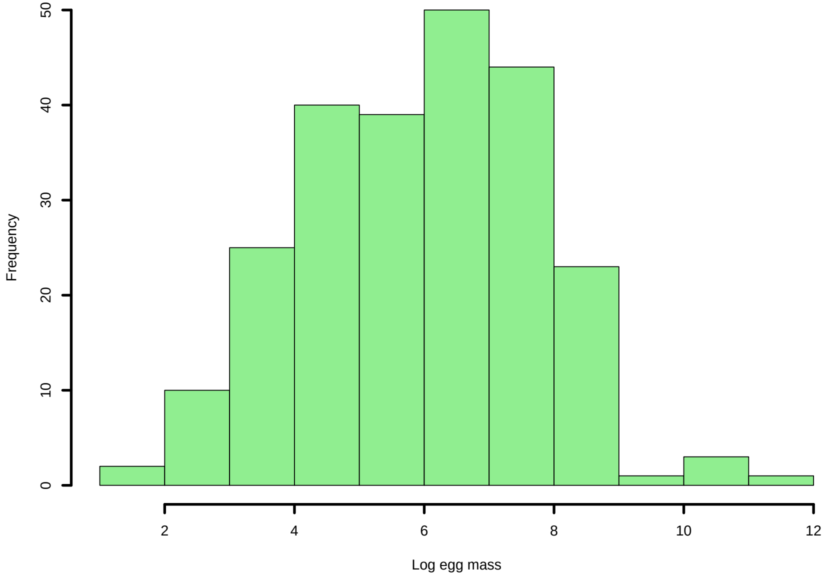
Normal Likelihood and Regression

We want to get to multiple regression, ANOVA etc. etc, but we will build up to it

Start with estimating the parmeters of a simple normal distribution

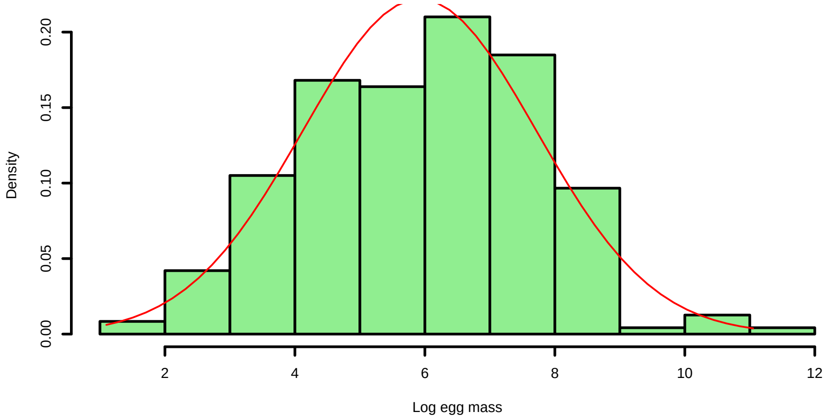
Normal Likelihoods

Some data: egg size in 239 species of bird



The Normal Distribution

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Our Task

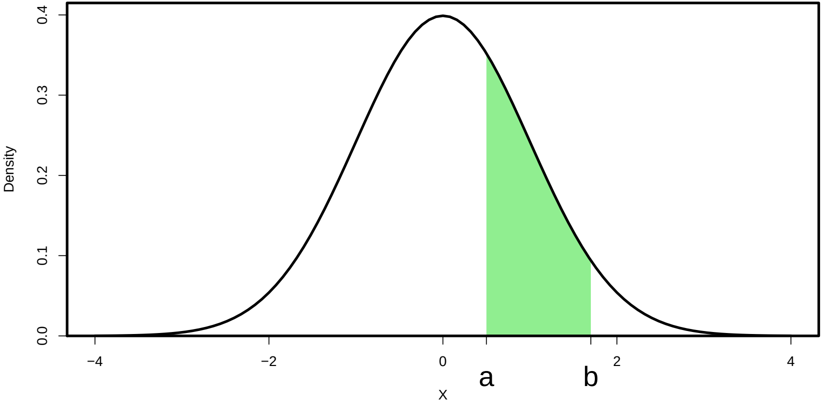
Estimate the parameters of this distribution - estimate $\hat{\mu}$ and $\hat{\sigma}^2$

These parameters tell us about how big eggs are, and how much variation there is - more complicated models have parameters, these parameters should say something useful scientifically

Normal Probabilities

The normal is continuous, so $Pr(X = x) = 0$, and we talk about probabilities of being between 2 values:

$$Pr(a < x < b|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_a^b e^{\frac{1}{2\sigma^2}(x-\mu)^2} dx$$



Why the Normal Distribution is Nice

The shape of the normal does not change with the parameters.

- Changing the mean moves the whole distribution
- Changing the variance stretches the distribution

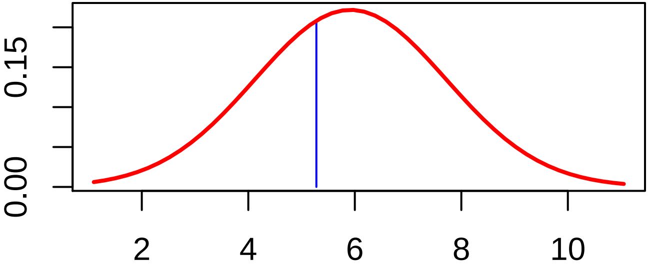
Lots of things look normal

- Central Limit Theorem

The Normal Likelihood

For one data point the likelihood is the probability density function ("pdf"):

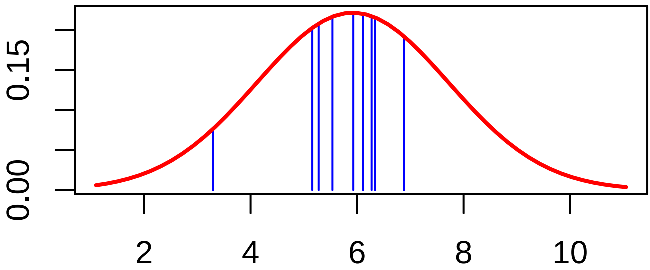
$$p(x_1|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1-\mu)^2}{2\sigma^2}}$$



The Normal Likelihood: lots of data

With many data poits, we assume they are independent (given the parameters). The likelihood is simply the product of each of the likelihoods:

$$p(x_1, x_2, x_3, ..., x_n|\mu, \sigma^2) = p(x_1)p(x_2)p(x_3)...p(x_n) = \prod_{i=1}^{n} p(x_i|\mu, \sigma^2)$$



The Full Normal Likelihood

Some notation: just like we use $\sum$ to mean a summation, we use $\prod$ to mean a product (i.e. multiplying terms together)

$$p(x_1, x_2, x_3, ..., x_n|\mu, \sigma^2) = \prod_{i=1}^{n} p(x_i|\mu, \sigma^2)$$

Plugging in the normal pdf we get

$$p(x_1, x_2, x_3, ..., x_n|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

i.e. scary

The Normal Likelihood

We can make this simpler by looking at the log-likelihood

$$l(\mathbf{x}|\mu, \sigma^2) = \sum_{i=1}^{n} l(x_i|\mu, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2\sigma^2}$$

and removing constants

$$l(\mathbf{x}|\mu, \sigma^2) = -\frac{n}{2}\log \sigma^2 - \sum_{i=1}^{n}\frac{(x_i - \mu)^2}{2\sigma^2}$$

So this is just quadratic in x_i

Maximising the likelihood

We want estimates of μ and σ^2, so we need to find values of μ and σ^2 which give the maximum of the likelihood

(Quick Break??)

Maximising the likelihood

There are two parameters: μ and σ^2

We can find the MLE of $\{\mu, \sigma^2\}$ by looking at these separately, i.e. solve $\frac{\partial l(\mathbf{x}|\mu,\sigma^2)}{\partial \mu}$ and $\frac{\partial l(\mathbf{x}|\mu,\sigma^2)}{\partial \sigma^2}$

This is mathematically more involved. We do it by finding $\hat{\mu}$, the the LME for μ:

$$
\begin{aligned}
l(\mathbf{x}|\mu, \sigma^2) &= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i^2 - 2x_i\mu + \mu^2) \\
&= -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)
\end{aligned}
$$

Differnetiate w.r.t (w.r.t = "with respect to") μ, set to zero, re-arrange, and pray

The Solutions

The MLE for the mean is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

For the variance, differentiate w.r.t σ^2, set to zero, re-arrange, pray some more, and get

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

For details, do it yourself or see
`https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood`

Comments

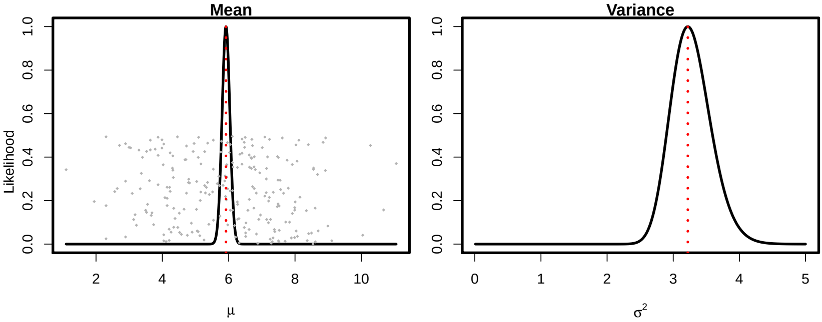
The estimate $\hat{\mu}$ is just the sample mean, and $\hat{\sigma}^2$ is just the sample variance

- the whole distribution can be summarised by these two statistics

$\hat{\sigma}^2$ has n as a denominator, not $n-1$

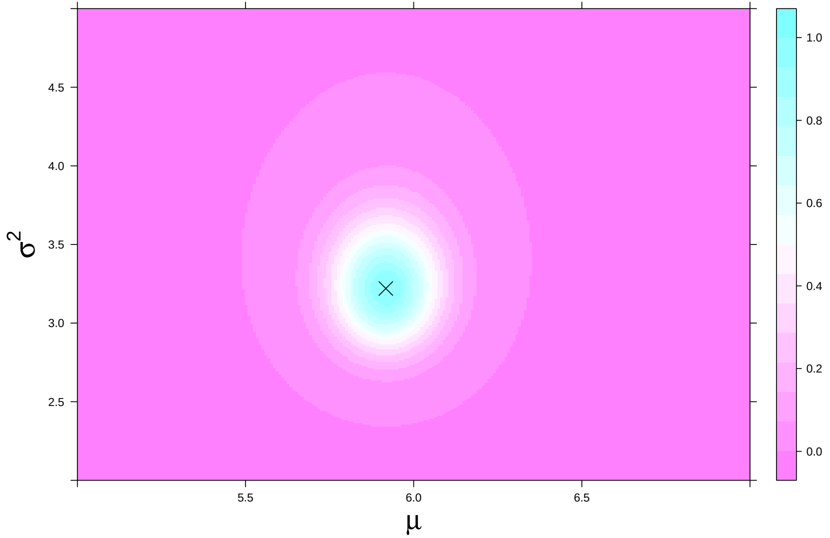
- because we assume the MLE for $\hat{\mu}$: using (n-1) is better because it takes into account the uncertainty

What it looks like



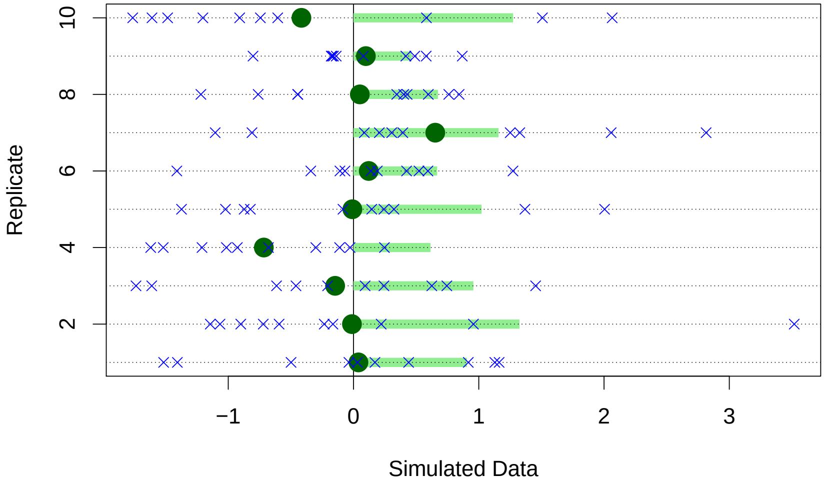
Joint Likelihood

The likelihoods for $\hat{\mu}$ and $\hat{\sigma}^2$ are independent



Uncertainty

Different realisations of the data give different results



We can summarise this with standard errors and confidence intervals

Back to maths

The first differential of the log-likelihood w.r.t. μ is

$$\frac{\delta l}{\delta \mu} = -\frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

If we differentiate again w.r.t $\hat{\mu}$ we get

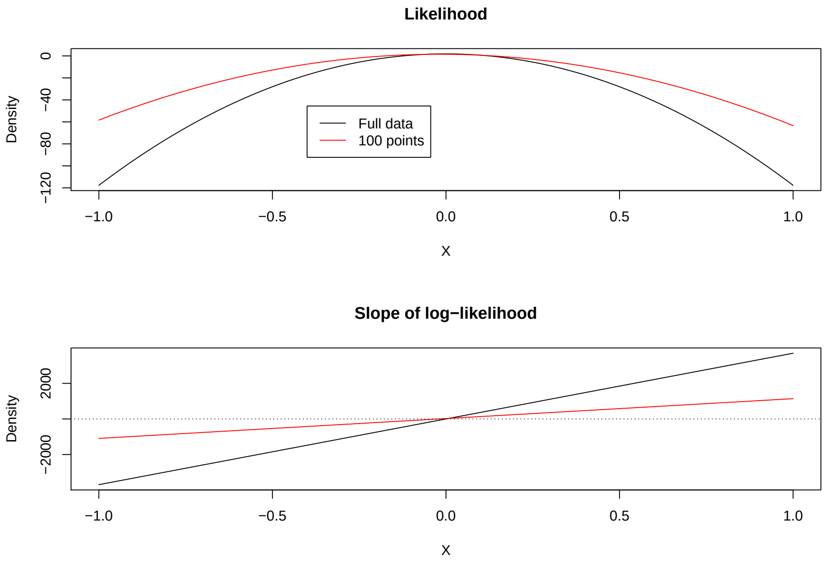
$$\frac{\delta^2 l}{\delta \mu^2} = -\frac{n}{\sigma^2}$$

The second differential is the curvature of the likelihood, and is the negative inverse of the standard error

This turns out to be a general assypmtotic approximation

- i.e. usually a good approximation, with enough data

Curves in Plots



Confidence Intervals for $\hat{\mu}$

$\hat{\mu}$ has a Normal likelihood with standard error $\hat{\sigma}^2/\sqrt{n}$

- proof not shown!

So we can use a Normal distribution to create the (asymptotic) ML confidence intervals

95% Confidence interval $= (\hat{\mu} - 1.96\sqrt{\hat{\sigma}^2/n}, \hat{\mu} + 1.96\sqrt{\hat{\sigma}^2/n})$

Confidence Intervals for $\hat{\sigma}^2$

The (asymptotic) standard error for $\hat{\sigma}^2$ is $\sqrt{2/n}\hat{\sigma}^2$, so the confidence interval is *approximately*

$$(\hat{\sigma}^2 - 1.96\sqrt{\frac{2}{n}}\hat{\sigma}^2, \hat{\sigma}^2 + 1.96\sqrt{\frac{2}{n}}\hat{\sigma}^2)$$

Why asymptotic?

The standard error is calculated at the MLEs of the other parameters

So any uncertainty in $\hat{\sigma}^2$ is ignored

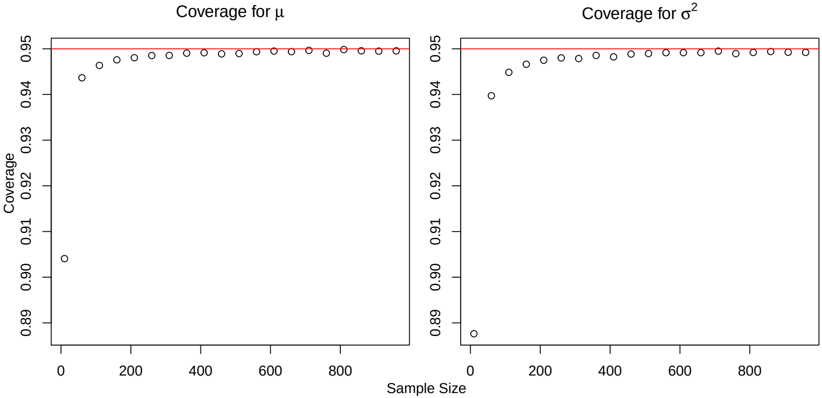
This is also why the confidence interval is normally distributed, not a t-distribution

When N is large, it is OK

For σ^2, the likelihood is not symmetric, but at reasonable sample sizes this does not matter

Coverage

The coverage is the proportion of times the confidence interval contains the true value. For a 95% confidence interval, this should be 95%.



Does the model fit?

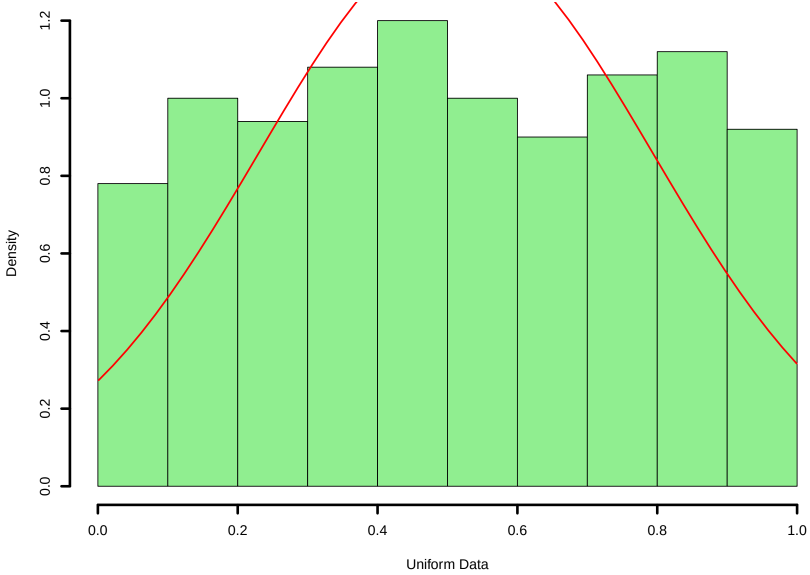
How could the model be wrong?

The assumptions:

- data follow a normal distribution
- data have a constant mean and variance
- data are independent

Does the model fit?

This is a simple model, but it could still fit badly. For example, this data does not look like a normal distribution



Checking Normality

We should check if our model is good, i.e. are the assumptions OK?

If not, we might have to change the model

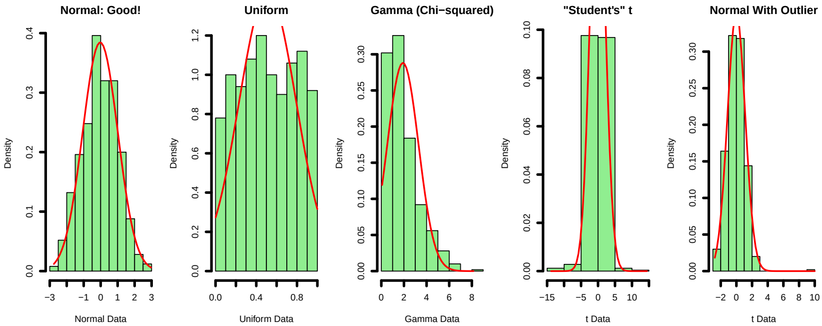
Checking Normality

For our bird egg data, the main question is whether the data are normally distributed

Later we will ask if we can improve the model by trying to explain the variation

Some Examples

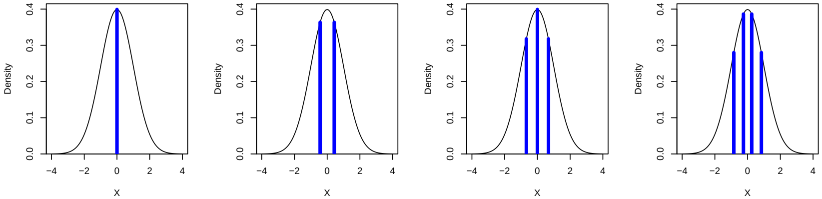
Some examples of good/bad fit



Quantile Plots

If we sample n points from some distribution, we can work out what the expected values are. One way to do this is to split the distribution up into equal slices, and put the expected data where the slices are

- one point: the expected point is the mean (or median!)
- two points: 33% & 67% quantiles
- three points: 25%, 50% (median) & 75% quantiles



Normal Probability Plots

With lots of points, the quantiles should be about the same as the (ordered) points

So, we can line them up

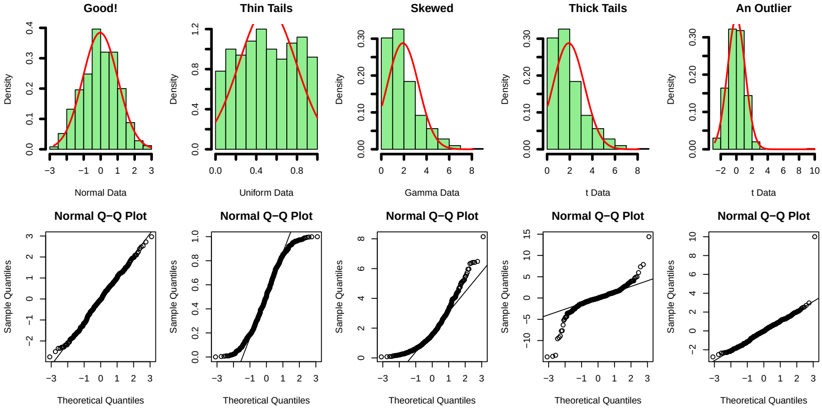
If we plot them against each other they should lie on the 1:1 line

Normal probability Plots

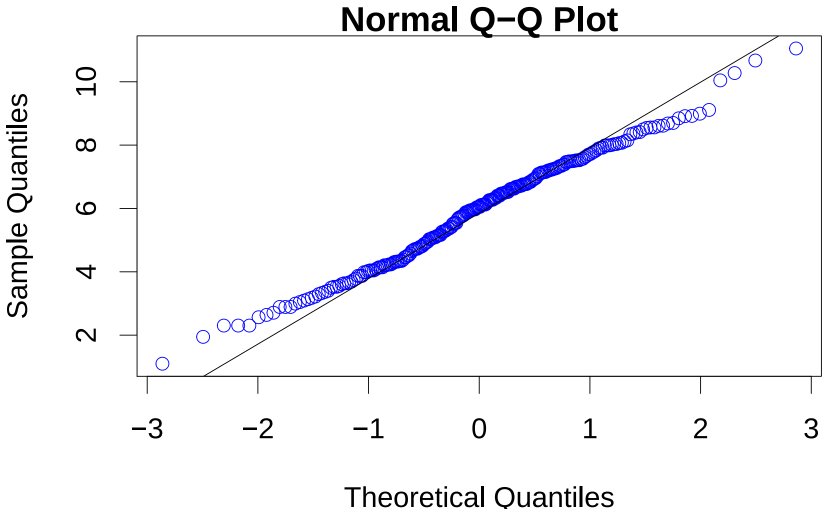
The normal distribution is nice because it has the same shape: we can move the x-axis or stretch it, but the shape is the same.

So we can plot the ordered data against the expected quantiles of a N(0,1) distribution with the same number of points

Normal probability Plots



Normal probability Plots



Suggests tails are too thin for normal, but not massively so

What we have Covered

The normal likelihood

- used a lot

Maximum likelihood with > 1 parameter

- maximumse w.r.t. each parameter

Standard errors and asymptotic confidence intervals

Model Checking

- does the model fit the data?

Next Week

Regression

Allowing μ to change

More modelling, less inference