

# Post-Graduate Statistics Course

Jarrold Hadfield

University of Edinburgh

## Lecture 1

Post-Graduate Statistics Course

Jarrold Hadfield  
University of Edinburgh

This is going to be an intensive 5 days and what I hope is that by the end you a) have a good understanding of the basic principles that underlie modern statistical inference and b) a thorough understanding of the basic models often encountered when dealing with biological data and c) an awareness of the common problems that are often met and how best to deal with them. Many of you will feel that you know the basics, and may get impatient to move on to the more advanced material, which we will cover in the later lectures.

- ANCOVA
- Bradley-Terry models
- MANCOVA
- Meta-analysis
- Multi-membership models
- Pedigree analysis: animal models
- Phylogenetic analysis: comparative approach
- Random Regression
- Rasch Models
- Regression
- Ridge Regression
- Splines
- Survival-analysis
- Threshold models
- Time-series

### Generalised Linear Mixed Models

Here's a shortlist of the type of advanced models we might cover. Some of these models I'm sure you're familiar with, some you'd probably like to be more familiar with, and there's some you've probably never heard of but they're useful nevertheless. It's a long shortlist, with names that are designed, or at least retained, because they can intimidate and scare the ignorant. However, the differences that exist between these models are absolutely tiny. They are all special cases of Generalised Linear Mixed Models. If you understand the basics - if you can fully understand a paired t-test - you can understand with little extra effort a Generalised Linear Mixed Model and therefore a whole suite of advanced statistical modelling techniques. So I urge you not to get impatient with the basics because they are the building blocks for everything else.

### └ Course Outline

Now it's going to be handy for me to know at what level people are and what types of analysis people are going to need. Probably the most useful thing to know is how many people are doing 'proper' experiments; randomised experiments with controls? You are the lucky ones, if you've been clever with your experimental design you don't need to be clever with your stats. How many people are working with non-experimental data? You are going to have to become good statisticians to make sense of your data. How many people are familiar with linear models? How many people are familiar with generalised linear models, so working with Poisson or binomial data? How many people have data like this but are not yet comfortable with analysing it? How many people know what random effects are and have used mixed models?

	Morning	Afternoon
Mon	The basics	
Tue		
Wed		
Thu		
Fri		

└ Course Outline

So this morning we're going to start with the fundamentals of statistical inference, and perhaps, if we have time we'll explore what a linear model is.

	Morning	Afternoon
Mon	The basics	
Tue		
Wed		
Thu		
Fri		

	Morning	Afternoon
Mon	The basics Linear Models	
Tue		
Wed		
Thu		
Fri		

└ Course Outline

Tomorrow we'll look a bit more into linear models

	Morning	Afternoon
Mon	The basics Linear Models	
Tue		
Wed		
Thu		
Fri		

	Morning	Afternoon
Mon	The basics	
Tue	Linear Models	
Wed	Generalised Linear Models	
Thu		
Fri		

└ Course Outline

and on Wednesday we'll move on to Generalised Linear Models. So these are models that are not only good for analysing continuous data that are close to being normally distributed, but also other types of data like count data.

	Morning	Afternoon
Mon	The basics	
Tue	Linear Models	
Wed	Generalised Linear Models	
Thu		
Fri		

### Course Outline

	Morning	Afternoon
<b>Mon</b>	The basics	
<b>Tue</b>	Linear Models	
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	
<b>Fri</b>		

	Morning	Afternoon
<b>Mon</b>	The basics	
<b>Tue</b>	Linear Models	
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	
<b>Fri</b>		

On Thursday we'll introduce something called mixed effect modelling,

### Course Outline

	Morning	Afternoon
<b>Mon</b>	The basics	
<b>Tue</b>	Linear Models	
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	
<b>Fri</b>	Mixed Models II	

	Morning	Afternoon
<b>Mon</b>	The basics	
<b>Tue</b>	Linear Models	
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	
<b>Fri</b>	Mixed Models II	

which we'll cover in more detail on Friday. It takes a long time to become proficient at these types of model, but as you saw in the previous slide, once you understand them you have a lot of statistical tools at your disposal.



### Course Outline

Each afternoon we'll do some computer practicals on some data sets I've collated and we'll start with some simple data from which we want to answer simple questions, and work up to more complex data and more complex questions.

	Morning	Afternoon
<b>Mon</b>	The basics	Simulation
<b>Tue</b>	Linear Models	Linear Model Fitting in R
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	Insights into Grumpiness
<b>Fri</b>	Mixed Models II	

	Morning	Afternoon
<b>Mon</b>	The basics	Simulation
<b>Tue</b>	Linear Models	Linear Model Fitting in R
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	Insights into Grumpiness
<b>Fri</b>	Mixed Models II	

### Course Outline

	Morning	Afternoon
<b>Mon</b>	The basics	Simulation
<b>Tue</b>	Linear Models	Linear Model Fitting in R
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	Insights into Grumpiness
<b>Fri</b>	Mixed Models II	Own data

	Morning	Afternoon
<b>Mon</b>	The basics	Simulation
<b>Tue</b>	Linear Models	Linear Model Fitting in R
<b>Wed</b>	Generalised Linear Models	
<b>Thu</b>	Mixed Models I	Insights into Grumpiness
<b>Fri</b>	Mixed Models II	Own data

On Friday afternoon, if you still have some energy, I'll let you try and put some of the things you've learnt into practice with your own data.

- What do we want to learn from the data?

### └ The basics

OK, today, we're going to start by asking a superficially easy question - what do we want to learn from the data we have collected?

- What do we want to learn from the data?
- Ingredients (Model, Parameters, Data)

### └ The basics

I want to persuade you to *model* your data; to construct a mathematical model of how you think the world works. We'll look at the primary ingredients that are needed; the model, the parameters of that model that we would like to learn about, and the data that we've collected in the hope that they will tell us something about these parameters and perhaps the plausibility of the model itself.

- What do we want to learn from the data?
- Ingredients (Model, Parameters, Data)

- What do we want to learn from the data?
- Ingredients (Model, Parameters, Data)
- Distributions (Data, Sampling, Posterior)

### └ The basics

We are going to have to understand a little bit about probability, and about probability distributions, whether that is the probability of our data (the data distribution), the probability of obtaining some parameter estimate (the sampling distribution) or the probability of some value being the true parameter value (the posterior distribution).

- What do we want to learn from the data?
- Ingredients (Model, Parameters, Data)
- Distributions (Data, Sampling, Posterior)

- What do we want to learn from the data?
- Ingredients (Model, Parameters, Data)
- Distributions (Data, Sampling, Posterior)
- Linear Model

### └ The basics

We are going to try and understand these concepts using data on grumpiness that people in IEB collected a few years ago, and using simple techniques such as t-tests and linear models.

- What do we want to learn from the data?
- Ingredients (Model, Parameters, Data)
- Distributions (Data, Sampling, Posterior)
- Linear Model

# What do we want to learn from the data?



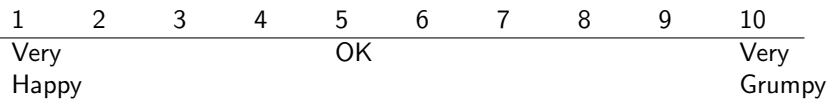
## Lecture 1

└ What do we want to learn from the data?

OK, let's start with some data we've already collected. I took two photos of people that worked in IEB - one when they were grumpy and one under 'normal' conditions. This is Matt Bell under normal conditions and Laura Ross when grumpy.

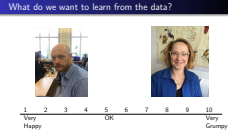


# What do we want to learn from the data?



## Lecture 1

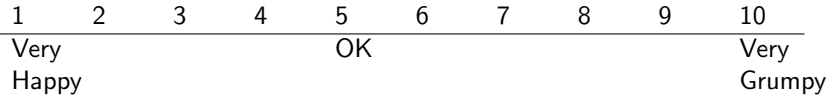
What do we want to learn from the data?



I then set up a survey in SurveyMonkey and asked people to assess the grumpiness of those photographed. 122 people took part. Now there's three types of question I could ask of these data.



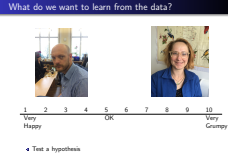
# What do we want to learn from the data?



- Test a hypothesis

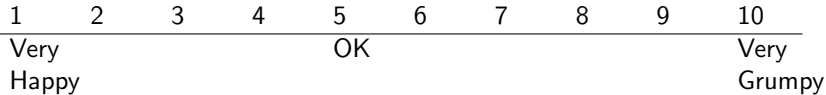
## Lecture 1

What do we want to learn from the data?



In the first instance I might just want to test a hypothesis: does someone in a grump look more grumpy? I don't care how much more grumpy they look, I just want to know do people look different when in a grump. This is the type of question many people are trained to answer - they do a statistical *test* and their eyes go straight to the p-value.

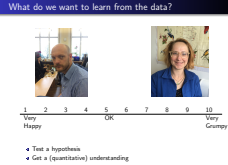
# What do we want to learn from the data?



- Test a hypothesis
- Get a (quantitative) understanding

## Lecture 1

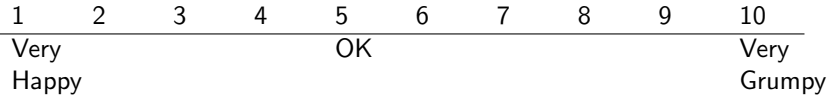
### What do we want to learn from the data?



We might want to ask a more nuanced question. How much more grumpy do people look when they're in a grump? Is it a big difference or is it a small difference? If our best answer is a big difference how confident are we in that answer? Now, there is a place for pure hypothesis testing in science, but I think that most scientists most of the time should be interested in a quantitative understanding. They should be using statistical *models* not statistical *tests*. That's not to say you shouldn't hypothesis test, but it should be done within the context of this richer form of inference. For example, if our best answer is there's a big difference but we were not very confident in it, I might then ask, are we at least confident that there is some difference - a hypothesis test.

If you are not working with experimental data, or you are, but the experimental design is not perfect, then you might need to put other things in the model in order to get a better answer. Does it matter that some of the respondents know the people being photographed? If someone is generally grumpy might we score their photo as grumpy even if they had a beaming smile? Would this affect our answer? When you become proficient at statistical modelling you will see that you can ask clever and informed questions of your data and get a much deeper insight into the biology of the problem than you could have, had you been restricted to simple hypothesis tests. It is obtaining a quantitative understanding of our problem that we're going to focus on in this course.

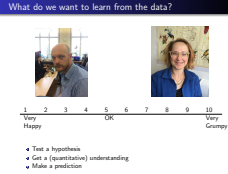
# What do we want to learn from the data?



- Test a hypothesis
- Get a (quantitative) understanding
- Make a prediction

## Lecture 1

### What do we want to learn from the data?



The final type of question you might ask of a body of data - and one we're not really going to focus on a lot - is prediction. What's the best way to use these data to predict whether someone is grumpy or not if we were given a new body data like the one we've previously collected? How best to use respondent's grumpiness scores to correctly classify people as being in a grump or not? There's a whole range of techniques - such as machine learning - that try and maximise the predictive power of the model. They're used by companies such as amazon and google to sell you crap you don't want or doesn't work, like degrumping pills. They have some utility in science but the problem is they don't tell you much about the biology of the problem. They can have a lot of complicated high-order terms that are hard to interpret, and for a scientist, I think a statistical model is there to extract meaning from the data in the simplest and most robust way possible.

So, we are really going to focus on this middle type of question. How can we model our data in a way that gives us a quantitative understanding of what is going on?

- Data
  - Response variable(s)
  - Predictor variable(s)

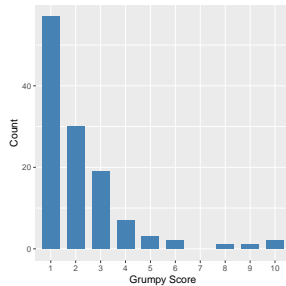
### Ingredients

To do this, the first - and by far the most important - ingredient is data, and generally we classify these into two types.

- Data
  - Response variable(s)
  - Predictor variable(s)

- Data

- Response variable(s)
- Predictor variable(s)



### Ingredients

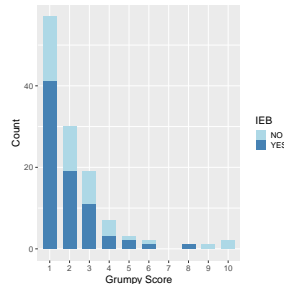
A response variable - the set of observations we are primarily trying to explain. So in our data this is the set of scores associated with our photographs. So these are the scores of the respondents that rated this photo of Laura. Most people thought Laura looked very happy, a few people, presumably with expressive agnosia, give her a 10 out 10 on the grumpy scale.

- Data
  - Response variable(s)
  - Predictor variable(s)



- Data

- Response variable(s)
- Predictor variable(s)



### Ingredients

We might also have some predictor variables that may explain some of the patterns we see in our response variable. These might be discrete predictors like whether the respondent is in IEB or not, or whether Laura is actually grumpy in this picture or not (this is grumpy Laura by the way). They might include continuous predictors such as age of the person photographed or the respondent. Sometimes these two types of data are given different names, such as dependent variable and independent variable but I always have to remind myself which is which. Some times predictor variables are called moderator variables or covariates, but I'll try and keep to response variable and predictor variables throughout.

- Data
  - Response variable(s)
  - Predictor variable(s)



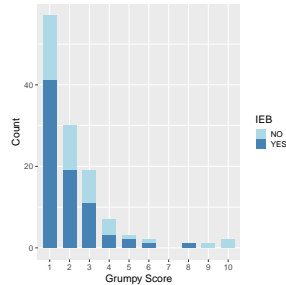
- Data

- Response variable(s)
- Predictor variable(s)



- Model


- What distribution do the data follow?
- How do the predictors change the data distribution?



### Ingredients

Ingredients

- Data
  - Response variable(s)
  - Predictor variable(s)
- Model
  - What distribution do the data follow?
  - How do the predictors change the data distribution?



The next thing we're going to need is a model - some idea about why these data look the way they do. We might start by specifying what type of distribution the response variable comes from. In this particular case it might actually be quite challenging - the data are on an ordinal scale that is bounded by 1 and 10 - so it might be quite hard to come up with a stochastic process that would generate data like this. We also need to make choices about which predictor variables to include in the model and make choices about how they affect this distribution.

- Data

- Response variable(s)
- Predictor variable(s)

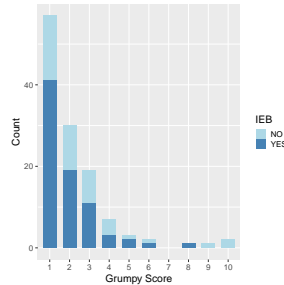


- Model

- What distribution do the data follow?
- How do the predictors change the data distribution?

- Parameters

- Location
- Dispersion




### Ingredients

Once we have a model it will be characterised by a number of parameters, and it's these parameters we want to learn something about. Often we can think about the parameters determining the location of the distribution - so for example we might want to consider a parameter that would allow non-IEB respondents to have a different average score than IEB respondents - or we can think about parameters determining the dispersion, so we might think the two types of respondent have the same average score but perhaps non-IEB respondents are more or less variable in their response.

Ingredients

- Data
  - Response variable(s)
  - Predictor variable(s)
- Model
  - What distribution do the data follow?
  - How do the predictors change the data distribution?
- Parameters
  - Location
  - Dispersion

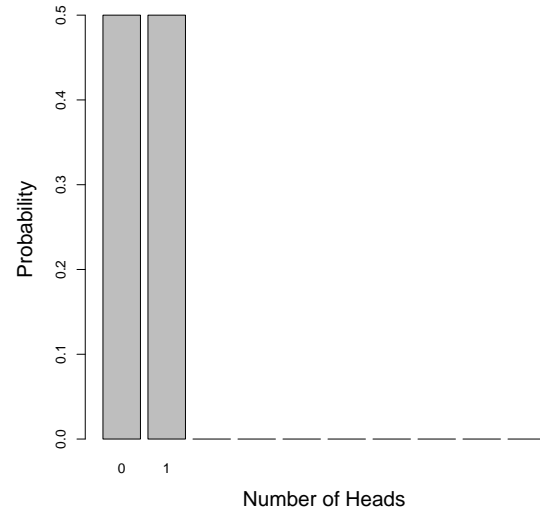
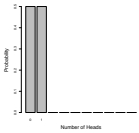




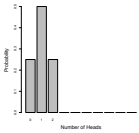
### └ Probability Distributions

In order to do statistics well you need to have a solid grasp of probability distributions and about uncertainty and randomness. This is a slippery subject and part of the confusion is that we use the word 'uncertainty' for two radically different phenomena. This is a 10p piece - as far as I can tell it's just a normal 10p piece. If I flip it what's the chance that it comes up heads? 50%.

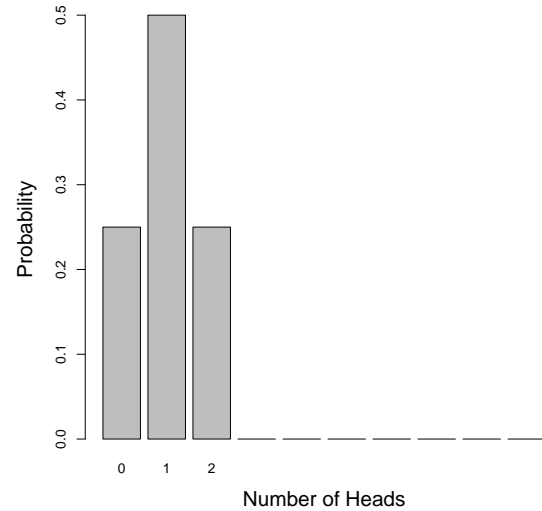
### Probability Distributions



This is a plot of a probability density function (or more accurately a probability mass function, since the outcomes are discrete). On the x-axis we have the possible outcomes; the number of heads, which for one flip of the coin can be either 0 or 1. On the y-axis we have the probability of getting that outcome. The chance of having no heads (a tail) is 50% and the chance of flipping a head is 50%. If I flipped the coin twice and counted the number of heads what would the probability mass function look like?

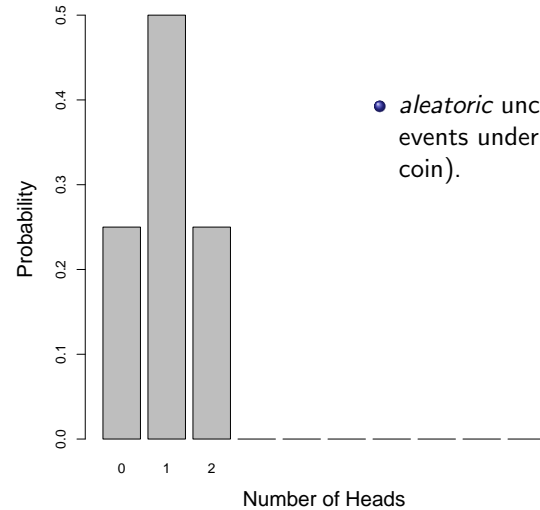
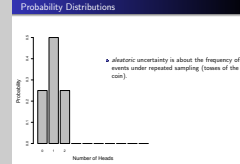


### Probability Distributions



The chance of no heads is 0.5 squared (25%). The chance of flipping a head then a tail is 25% and the chance of flipping a tail then a head is 25% so together the chance of one head is 50%. Finally, the chance of flipping two heads is 0.5 squared (25%).

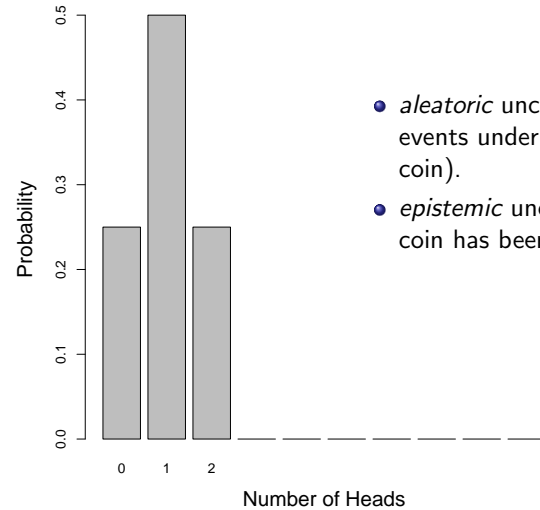
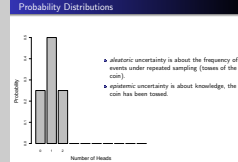
### Probability Distributions



- *aleatoric* uncertainty is about the frequency of events under repeated sampling (tosses of the coin).

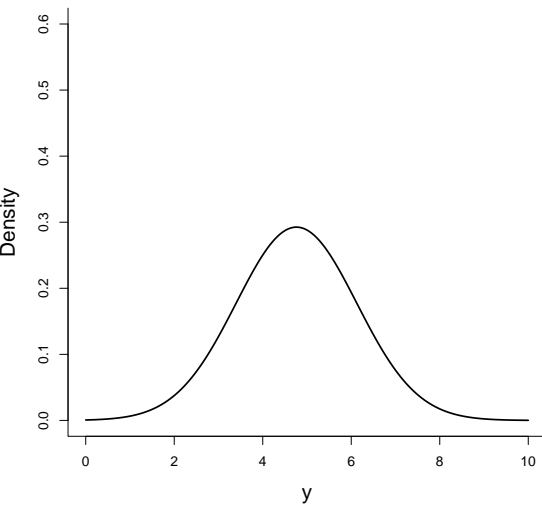
This type of uncertainty is known as *aleatoric* uncertainty; it is about the frequency of events under repeated sampling (tosses of the coin). Now let's say I had flipped the two coins and I can see the outcome but you can't. What is the chance that I've flipped one head and one head only? 50% again, right? But, this number 50% is referring to something completely different than it did before. The coins have been tossed; there is either one head or there is not with probability 1. The value of 50% that you are quoting is about uncertainty in your knowledge.

### Probability Distributions

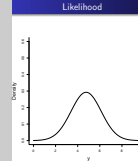


- *aleatoric* uncertainty is about the frequency of events under repeated sampling (tosses of the coin).
- *epistemic* uncertainty is about knowledge, the coin has been tossed.

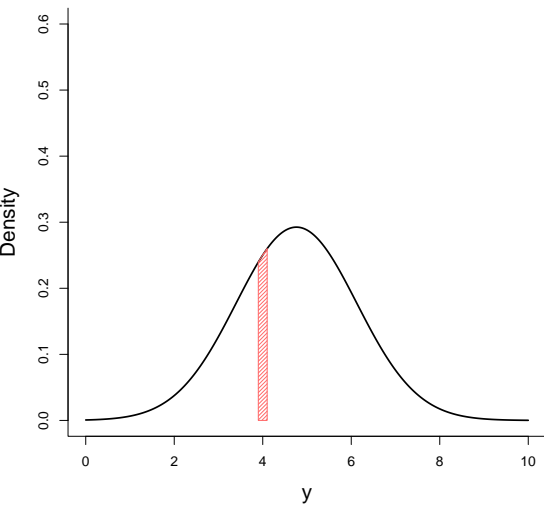
It's called epistemic uncertainty; for you the probability is 50% that there is one head but for me the probability is 100% that there is 2 heads. And now the probability that there is one head has changed for you too: it's 0%.



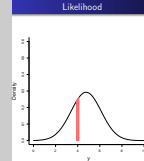
## Likelihood



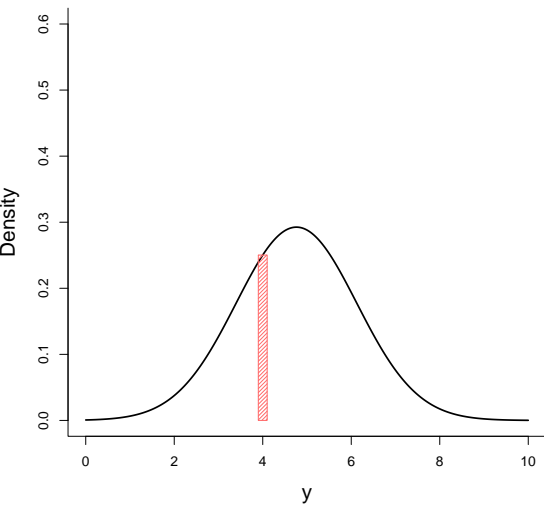
When flipping a coin the outcome is discrete and so the function that describes the probability of each outcome is called a probability mass function. This is the equivalent function for a continuous outcome that has a normal (or Gaussian) distribution - I'm sure your familiar with the probability density function of the normal. The chance of getting a specific value under the normal distribution, let's say 0.373244282955052, is? Well, it's vanishingly small. So you can't just read off a probability on the y-axis and find out the probability of an outcome like we did with the probability mass function we saw previously. However, relative differences are meaningful, so the probability density at a value of 4 is about 0.25 and the probability density at a value of 2 is about 0.04 so a value of 4 is about 6-7 times more likely than a value of 2.



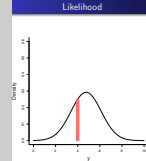
## Likelihood



Perhaps an easier way for you to think about a probability density function is to chop it up into little bits and think about it as a probability mass function. So for example, the probability of  $y$  being between 3.9 and 4.1 is equal to the area under the curve.

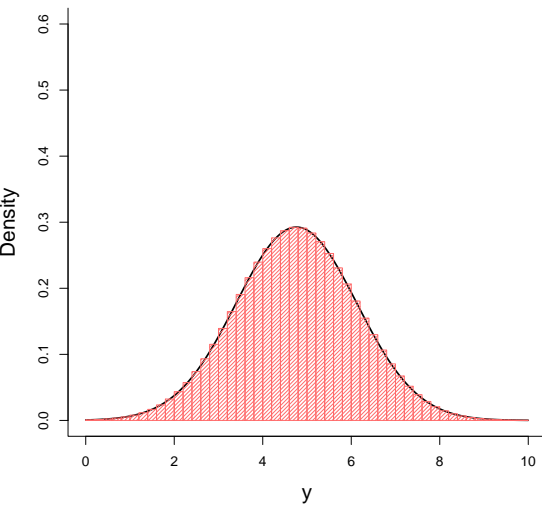


Likelihood

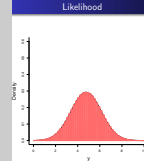


And this area is about 0.05, so there's a 5% chance that a value drawn from this distribution is between 3.9 and 4.1. We could represent this area as a rectangle.

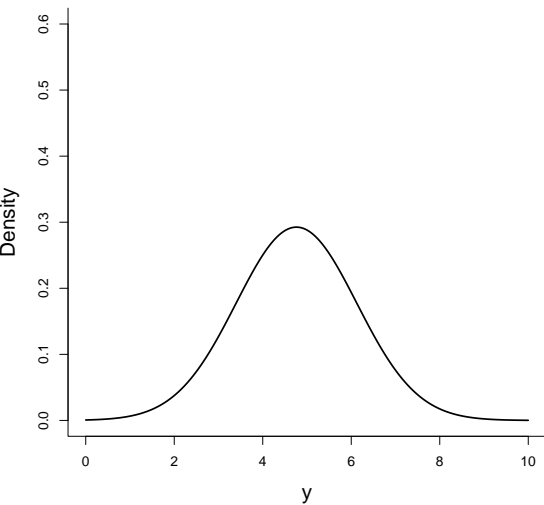




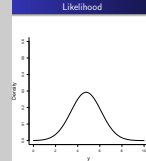
## Likelihood



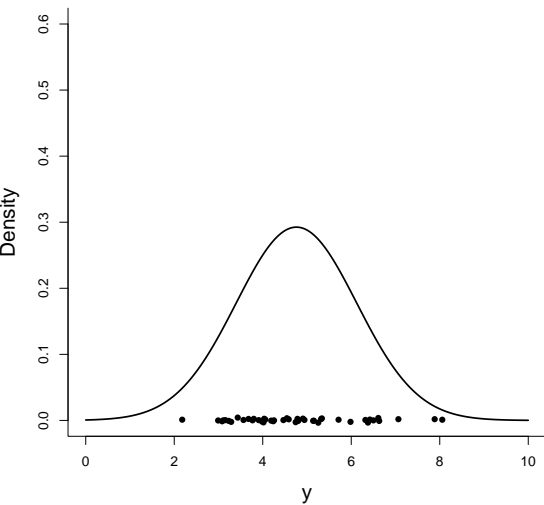
And if we liked we could also calculate the probability of lying in other equally spaced intervals so we had something that looked like a histogram. As with the probability mass function the area under this curve is equal to one;  $y$  must take some value with probability one. As we make these intervals smaller and smaller the histogram starts to get closer and closer to the probability density function, and is exactly equal to it when the interval is an infinitesimally small amount (often denoted  $dx$ )



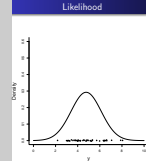
Likelihood



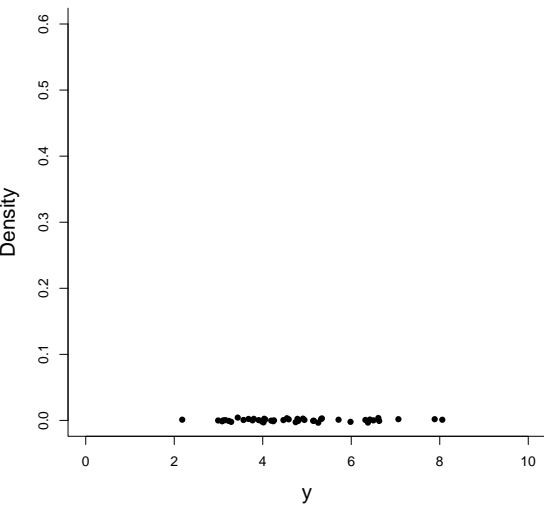
Ok, so this is our normal distribution - it has a mean of around 4.8 and a variance 1.9.



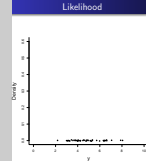
### Likelihood



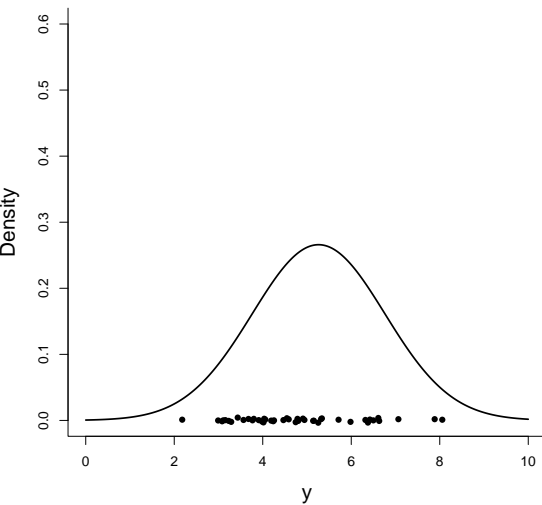
Lets say we observed some data that were generated from this distribution.



## Likelihood



Now of course the difficult thing is we don't know what distribution our data came from. We just have the data and we would like to use these data to try and understand something about the process that gave rise to them. One possibility would be to make a guess

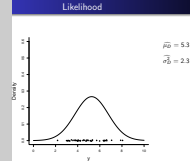


$$\hat{\mu}_D = 5.3$$

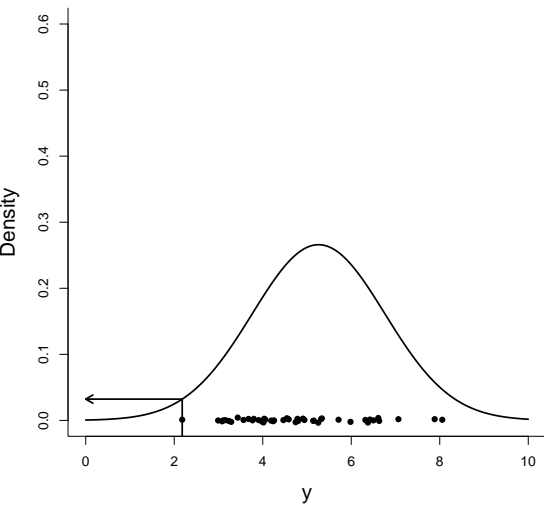
$$\hat{\sigma}_D^2 = 2.3$$



## Likelihood



So let's say we think it could be a normal distribution with a mean of 5.3 and a variance of 2.3. It is worth spending a bit of time on the notation.  $\mu$  and particularly  $\sigma^2$  are standard notation for the mean and variance. I have subscripted these quantities with a  $D$  to indicate they are the parameters of the *data* distribution. It is very important that you realise that these are parameters of the underlying distribution (often called population parameters):  $\mu_D$  is not the mean of the data you have observed (the sample mean) but the true underlying mean. The hat is standard notation for 'an estimate of' and so  $\hat{\mu}_D$  is an estimate of the true underlying population mean. So how plausible is a mean of 5.3 and a variance of 2.3? Well what we could do is ask what is the probability of observing these data if this were true: what is the *likelihood* of the data under this scenario?

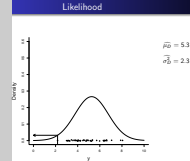


$$\hat{\mu}_D = 5.3$$

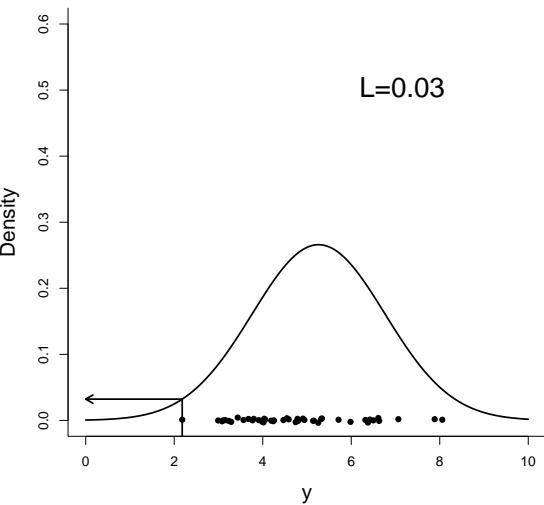
$$\hat{\sigma}_D^2 = 2.3$$



## Likelihood



As we saw earlier the chance of obtaining exactly these values is vanishingly small, but if we want to make relative statements then we can use the probability density. So we can read off the probability density for our smallest observed value ( $y=2.2$ )

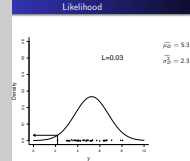


$$\hat{\mu}_D = 5.3$$

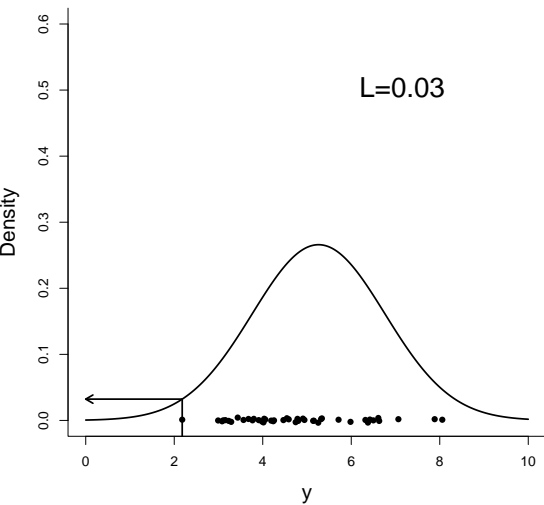
$$\hat{\sigma}_D^2 = 2.3$$

L

Likelihood



which is about 0.03. When we are talking about the probability of our data under some model we usually use the word likelihood rather than probability.



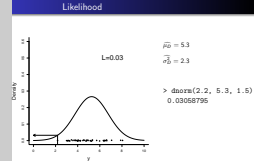
$$\hat{\mu}_D = 5.3$$

$$\hat{\sigma}_D^2 = 2.3$$

```
> dnorm(2.2, 5.3, 1.5)
0.03058795
```

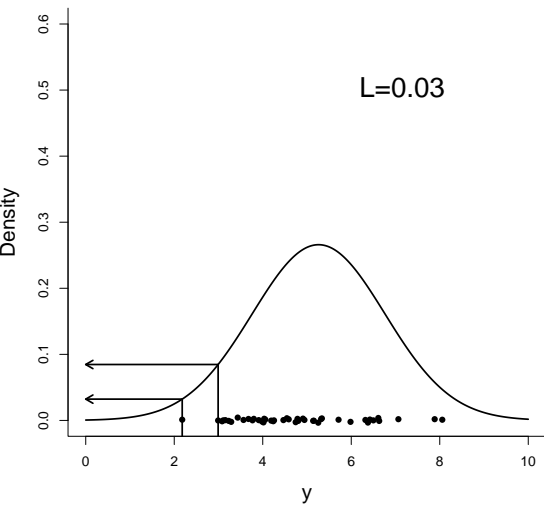


## Likelihood



you can calculate this in R using the `dnorm` function. The first argument is the value for which you want to calculate the density, the second argument is the mean of the distribution and the third argument is the standard deviation of the distribution. The standard deviation is the square root of the variance so in this case `sqrt(2.3)=1.5`.





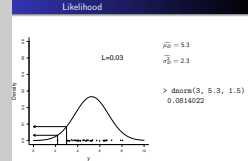
$$\hat{\mu}_D = 5.3$$

$$\hat{\sigma}_D^2 = 2.3$$

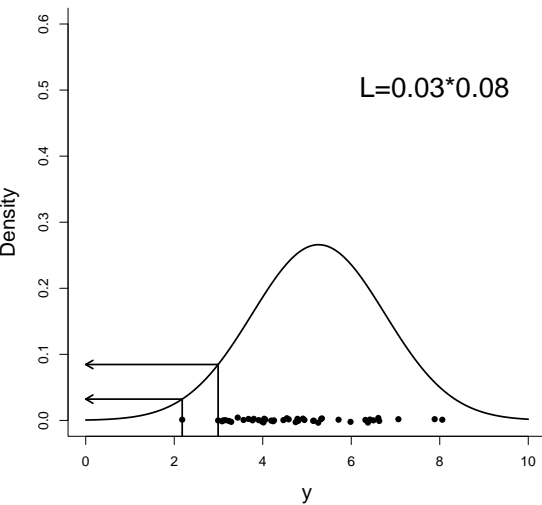
```
> dnorm(3, 5.3, 1.5)
0.0814022
```



Likelihood



we could then go to the next value ( $y=2.99$ ) and calculate its density which is about 0.08



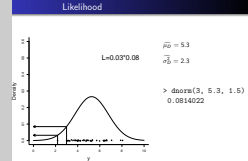
$$\hat{\mu}_D = 5.3$$

$$\hat{\sigma}_D^2 = 2.3$$

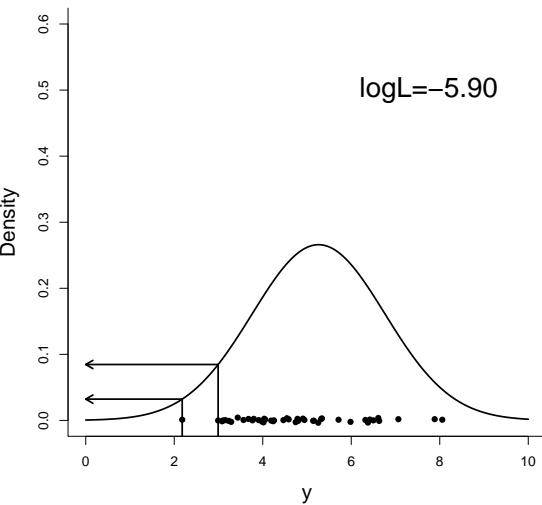
```
> dnorm(3, 5.3, 1.5)
0.0814022
```



Likelihood



and we can multiply these two numbers together. We are assuming that these data are drawn independently from this distribution and so we can multiply their probabilities together to get the probability of observing both these values. As the number of data points grows this product starts to become very small and so



$\log L = -5.90$

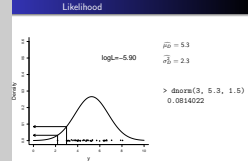
$$\hat{\mu}_D = 5.3$$

$$\hat{\sigma}_D^2 = 2.3$$

```
> dnorm(3, 5.3, 1.5)
0.0814022
```

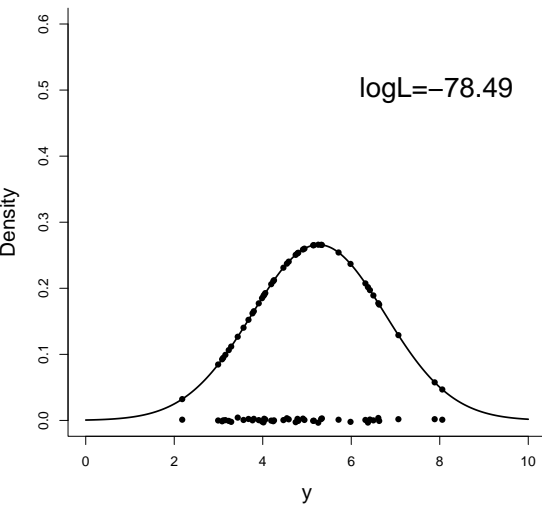


## Likelihood



it's easier to work with the log of the likelihood<sup>[1]</sup>.

<sup>[1]</sup> The log transformation is a monotonic transformation which means that if we increase  $x$ ,  $\log(x)$  also increases, and if we reduce  $x$ ,  $\log(x)$  also decreases. This means that values for which  $x$  is maximised (or minimised) are the same values for which  $\log(x)$  is maximised (or minimised) and so the values that maximise the likelihood are the same values that maximise the log-likelihood. However, the log-likelihood is easier to work with mathematically.



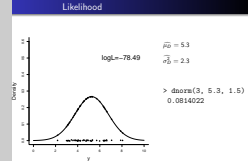
$$\hat{\mu}_D = 5.3$$

$$\hat{\sigma}_D^2 = 2.3$$

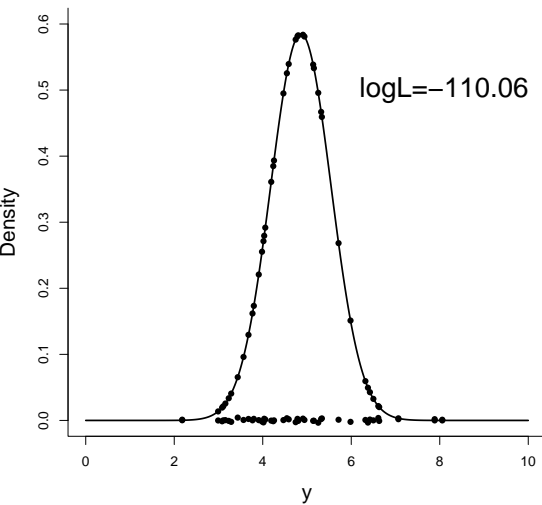
```
> dnorm(3, 5.3, 1.5)
0.0814022
```



## Likelihood



We can then work our way through all the data, reading off the height of the probability density function in each case, and we get the log-likelihood of observing these data given this distribution (a value of -78.49). We could then try different distributions to see if the likelihood of observing the data increases or decreases.

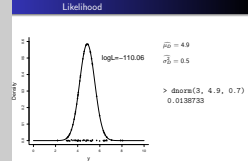


$$\hat{\mu}_D = 4.9$$

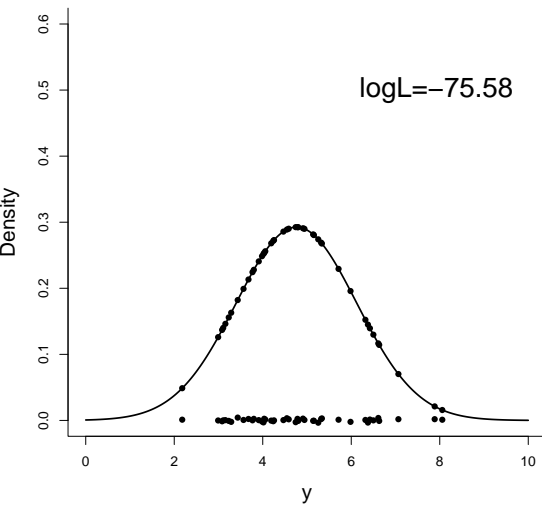
$$\hat{\sigma}_D^2 = 0.5$$

```
> dnorm(3, 4.9, 0.7)
0.0138733
```

## Maximum Likelihood



So we could shift the mean slightly to the left (4.9) and we could make the variance considerably smaller (0.5). We can see that under this distribution the log-likelihood, and therefore the likelihood, of the data has gone down. It's true that the observations close to the mean are more probable under this new distribution, but the values further from the mean are very very unlikely to have arisen given a variance of only 0.5. For example, the probability of observing our second smallest value of y has gone down by approximately a factor of 6. So in aggregate these data are less probable under this distribution than the one that preceded it.

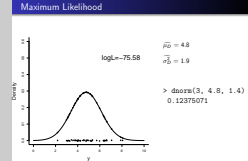


$$\hat{\mu}_D = 4.8$$

$$\hat{\sigma}_D^2 = 1.9$$

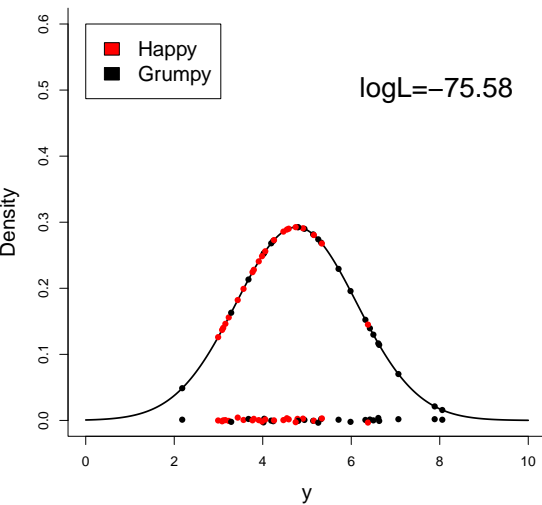
```
> dnorm(3, 4.8, 1.4)
0.12375071
```

### Maximum Likelihood



The techniques that we will use in this course use algorithms that can find the parameters (in this case the mean and variance) that maximise the likelihood of the data. In this simple case the algorithm is super simple and the parameters that maximise the likelihood are a mean of 4.8 and a variance of 1.9<sup>[1]</sup>. But these algorithms can be very general.

<sup>[1]</sup> In fact this is not the maximum likelihood (ML) estimator but the restricted maximum likelihood (REML) estimator, but more on this later ....

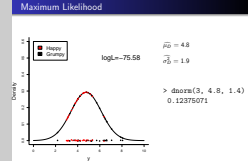


$$\hat{\mu}_D = 4.8$$

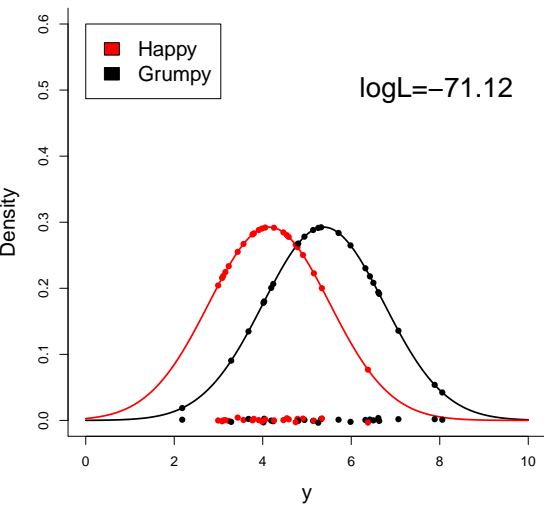
$$\hat{\sigma}_D^2 = 1.9$$

```
> dnorm(3, 4.8, 1.4)
0.12375071
```

### Maximum Likelihood



So for example, lets say these observations were one of two types; a grumpy type and a happy type. In fact, these are the grumpiness scores (averaged over respondents) for each of the 44 photos scored.



$$\hat{\mu}_D = 5.4$$

$$\hat{\sigma}_D^2 = 1.9$$

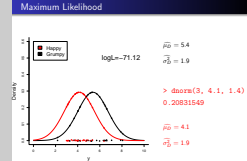
```
> dnorm(3, 4.1, 1.4)
0.20831549
```

$$\hat{\mu}_D = 4.1$$

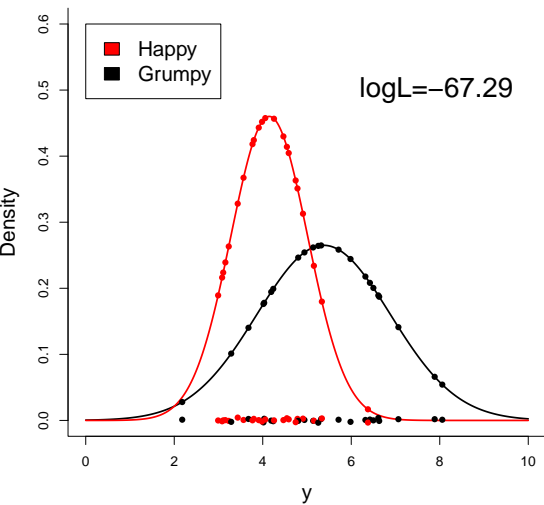
$$\hat{\sigma}_D^2 = 1.9$$

### Maximum Likelihood

We could imagine wanting three parameters where the mean of the grumpy and happy photos can differ (but their variance is assumed the same) and maximise the likelihood under this new scenario. We can see that the data are more probable if the mean of happy photos is less than that of grumpy photos. For example, the probability of our second smallest score, which was for a photo taken under happy conditions, has nearly doubled.







$$\hat{\mu}_D = 5.4$$

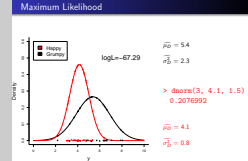
$$\hat{\sigma}_D^2 = 2.3$$

```
> dnorm(3, 4.1, 1.5)
0.2076992
```

$$\hat{\mu}_D = 4.1$$

$$\hat{\sigma}_D^2 = 0.8$$

### Maximum Likelihood



We could also allow the grumpy and happy photos to have different variances, and in this case the variance of the happy scores is estimated to be considerably less than the grumpy scores. You will notice that every time I added a parameter the likelihood has gone up and never went down. This has to be true; each new parameter introduces a little more flexibility that allows the distribution to capture more of the patterns in the data.

Now what type of uncertainty do you think the likelihood is dealing with? Aleatoric or Epistemic?

Aleatoric! There is no epistemic uncertainty; the data have been collected, I've plotted it and we can all see it.

# Posterior Distribution

*Likelihood*: the *aleatoric* probability of the data *given* a parameter value.

## Lecture 1

### └ Posterior Distribution

The likelihood is about aleatoric uncertainty. It is the probability of observing these (known) outcomes had they been generated under the particular distribution we are considering. Now you might wonder why on earth you would be interested in the likelihood. Why would I care about the probability of my data? I already have my data - whether the probability of those data is 0.01 or 0.5 is largely irrelevant - these are the data I have. It's not the probability of the data I'm interested in but the process that gave rise to them.

# Posterior Distribution

*Likelihood*: the *aleatoric* probability of the data *given* a parameter value.

*Posterior Distribution*: characterises *epistemic* uncertainty about the *true* parameter value.

## Lecture 1

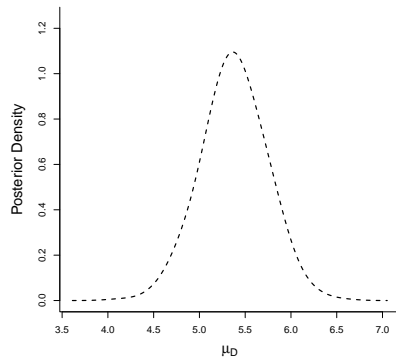
### └ Posterior Distribution

The uncertainty you want to understand is the epistemic uncertainty about the parameter values of your model. There's no aleatoric uncertainty about a parameter, there's a single underlying true parameter value, and the uncertainty arises because you cannot directly observe it and only have partial information about it through the data you observe. The distribution that characterises your epistemic uncertainty about a parameter value is called a posterior distribution.

# Posterior Distribution

*Likelihood*: the *aleatoric* probability of the data *given* a parameter value.

*Posterior Distribution*: characterises *epistemic* uncertainty about the *true* parameter value.



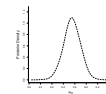
## Lecture 1

### Posterior Distribution

#### Posterior Distribution

*Likelihood*: the aleatoric probability of the data given a parameter value.

*Posterior Distribution*: characterises epistemic uncertainty about the true parameter value.

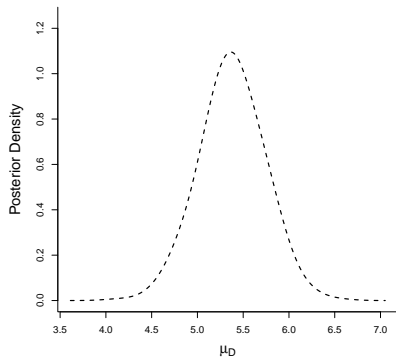


This is the posterior distribution for the mean score for grumpy photos. We can see that the most probable value of the true underlying mean is around 5.4. Values of around 5 or 6 are about half as probable as this, and most likely the true value lies between 4.5 and 6.5. The posterior distribution has a very easy straightforward interpretation. But the issue is that

# Posterior Distribution

*Likelihood*: the *aleatoric* probability of the data *given* a parameter value.

*Posterior Distribution*: characterises *epistemic* uncertainty about the *true* parameter value.



$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

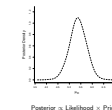
## Lecture 1

### Posterior Distribution

#### Posterior Distribution

*Likelihood*: the aleatoric probability of the data given a parameter value.

*Posterior Distribution*: characterises epistemic uncertainty about the true parameter value.



the posterior distribution is proportional to ( $\propto$ ) the likelihood (see it is useful!) multiplied by something called a prior. This is Bayes Theorem. The prior represents your state of knowledge (your epistemic uncertainty) about the mean score of grumpy photos before you'd observed any scores. Everybody loves the posterior distribution, most people hate the prior distribution; it might be hard to accurately convey your prior epistemic uncertainty about the true value, and if you get it wrong maybe the posterior distribution will be dominated by this error.

In some cases people are comfortable with using a prior. When I asked you what was the chance of flipping a head from this coin you said 50% because you have a strong prior belief that coins are fair. Let's say I flipped the coin 4 times and got 1 head and 3 tails. If I wanted to calculate the maximum likelihood estimate of the probability of flipping a head it would be 1 divided by 4 (0.25). Do you think this is the best estimate? No, you would still prefer an estimate of 0.5, I think.

I'm not going to cover Bayesian statistics in this course. I'm going to cover Frequentist methods; methods that attempt to say something about the plausibility of a particular parameter value from the likelihood alone.

*Sampling distribution:* characterises *aleatoric* uncertainty about *estimates*.

### └ Sampling Distribution

To understand how we can make inferences from the data alone - from the likelihood - we need to understand something called a sampling distribution. Earlier we used maximum likelihood to estimate the mean of the distribution from which we assumed our grumpy scores were drawn. Now imagine that you redid the experiment keeping the sample sizes the same. There's a few ways you could do it. You could use the same photos but get a new set of respondents. You could get some new photos of the same people and use the same set of respondents. The way in which you repeat the experiment would reflect what you were interested in learning. Lets say we repeated it using new photos of different people using a new set of respondents and from the new data we obtained a second estimate of the mean. Then repeat the experiment again and obtain a third estimate, and then a fourth estimate and so on. The resulting distribution of estimates is the sampling distribution. Of course, in reality people don't generate the sampling distribution empirically by repeating the same experiment over and over again. What they do is work out what it would be under certain assumptions.

*Sampling distribution:* characterises *aleatoric* uncertainty about *estimates*.

Mind-bending ... but often similar to a posterior distribution.

### └ Sampling Distribution

Now it requires a lot of mental gymnastics to understand what sampling distributions tell us about the world, but I think it is important to realise that they are often similar to posterior distributions and they are often interpreted informally as posterior distributions by scientists and even applied statisticians. I think this is OK.

*Sampling distribution:* characterises *aleatoric* uncertainty about *estimates*.

Mind-bending ... but often similar to a posterior distribution.

If you are a scientist rather than a statistician I want you to deceive yourselves that the sampling distribution is a posterior distribution but at the same time I want you to keep it in the back of your minds that you're being deceitful.

### Sampling Distribution

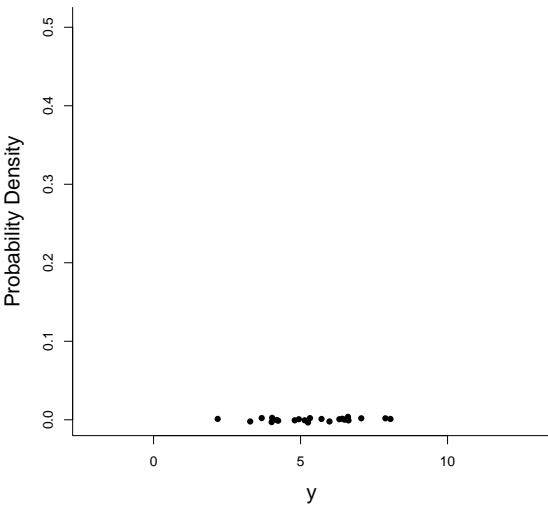
*Sampling distribution:* characterises *aleatoric* uncertainty about *estimates*.

Mind-bending ... but often similar to a posterior distribution.

If you are a scientist rather than a statistician I want you to deceive yourselves that the sampling distribution is a posterior distribution but at the same time I want you to keep it in the back of your minds that you're being deceitful.

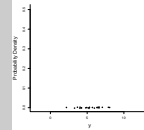
If you are a scientist rather than a statistician I want you to deceive yourselves that the sampling distribution is a posterior distribution but at the same time I want you to keep it in the back of your minds that you're being deceitful. You've got more important things to do than disappear down this rabbit hole.





Sampling Distribution

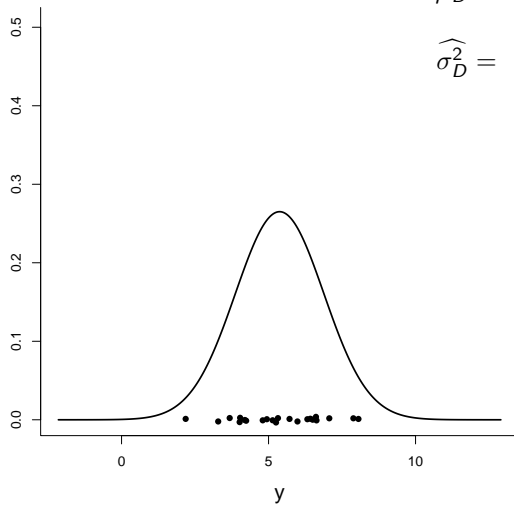
OK - these are the mean scores for our 22 grumpy photos



$$\widehat{\mu}_D = 5.38$$

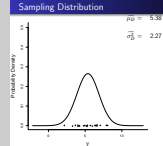
$$\widehat{\sigma}_D^2 = 2.27$$

Probability Density



### Sampling Distribution

and these are our maximum likelihood estimates of the mean and variance. For a simple problem such as this, the maximum likelihood estimate of the true underlying mean is the mean of the data ( $\widehat{\mu}_D = \bar{y}$ ) but it is important to realise these two quantities are fundamentally different things:  $\bar{y}$  is a function of the data (a statistic, the sample mean in this case) and  $\widehat{\mu}_D$  is an estimate of the population parameter  $\mu_D$  - the true mean.

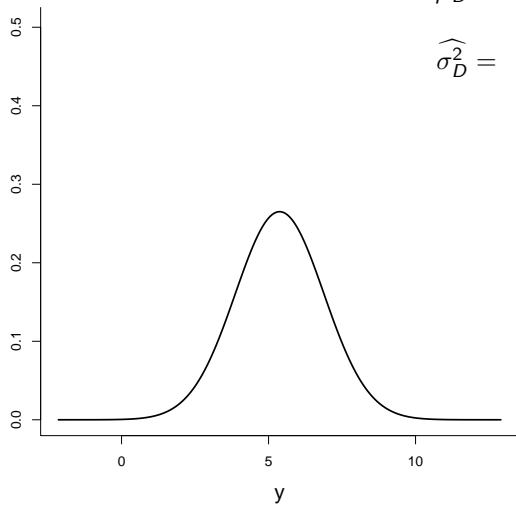


# Sampling Distribution

$$\widehat{\mu}_D = 5.38$$

$$\widehat{\sigma}_D^2 = 2.27$$

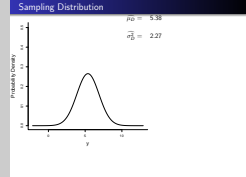
Probability Density



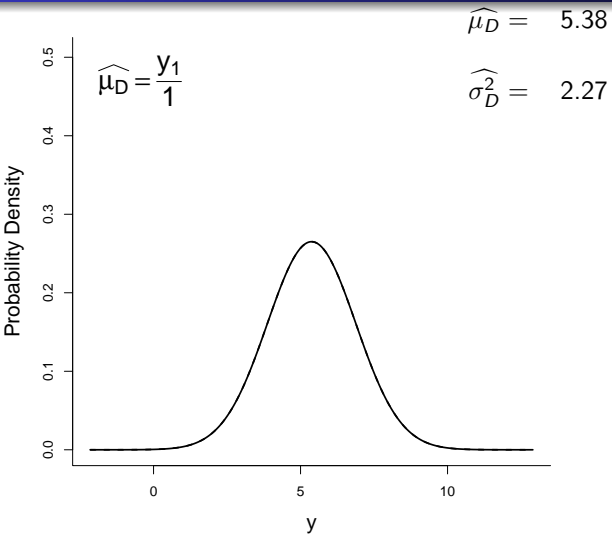
## Lecture 1

└ Sampling Distribution

Now let's discard the data and assume that our maximum likelihood estimate of the distribution is actually the truth.

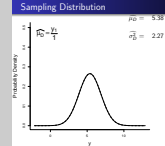


# Sampling Distribution



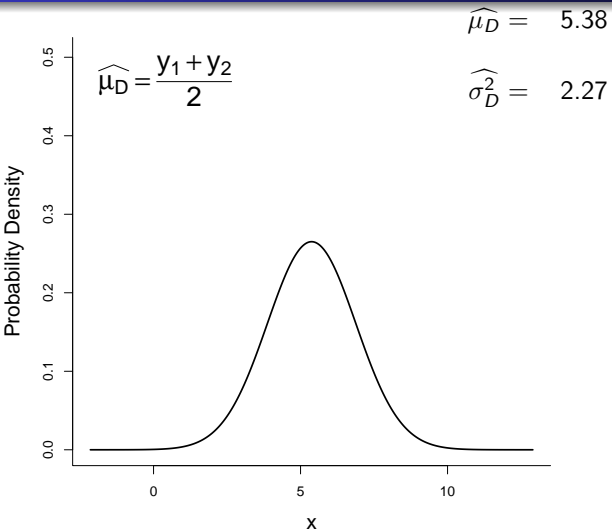
## Lecture 1

### Sampling Distribution



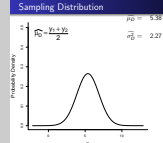
Lets also imagine that we were super lazy and we could only be bothered to collect one data point. The best estimate of the mean  $\widehat{\mu}_D$  is to take the mean of our observations (the sample mean) and so our best estimate is to simply use the number for the single data point we have. If we made a million lazy people rerun the experiment, each only collecting one data point and getting their own estimate what would the distribution of the estimates - the sampling distribution - look like? It would be exactly equal to the data distribution. We are after all just pulling single numbers from the data distribution.

# Sampling Distribution



## Lecture 1

### Sampling Distribution



Let's imagine we were now a bit less lazy and collected two observations ( $y_1$  and  $y_2$ ). Our best estimate of the underlying mean is again the sample mean  $(y_1 + y_2)/2$ . If we kept pulling pairs of observations from our data distribution (the black line), took their average, and then plotted their distribution (the sampling distribution) what do you think it would look like?

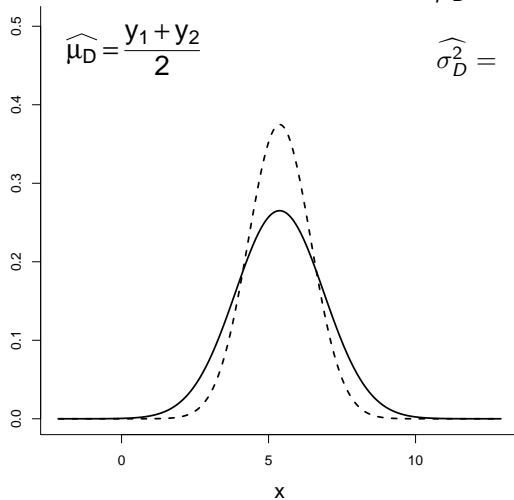
# Sampling Distribution

$$\widehat{\mu}_D = 5.38$$

$$\widehat{\sigma}_D^2 = 2.27$$

$$\widehat{\mu}_D = \frac{y_1 + y_2}{2}$$

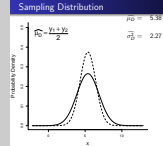
Probability Density



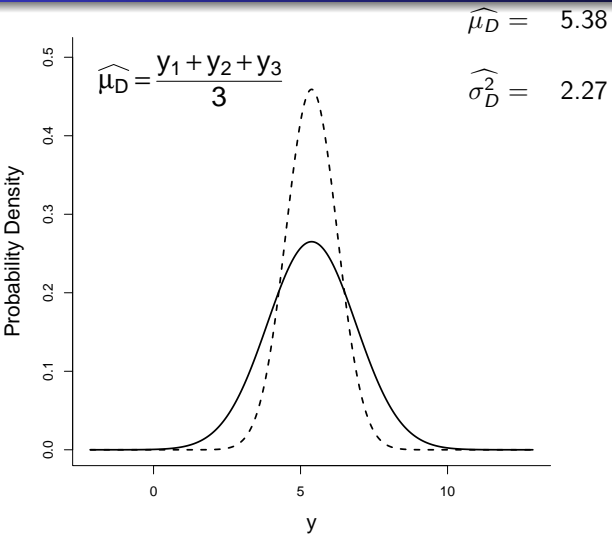
## Lecture 1

### Sampling Distribution

The first thing to notice is that the mean of the sampling distribution is equal to the true underlying mean. The second thing to notice is that the width of our sampling distribution is now narrower than what it was before when we only had a single observation. This makes sense. A value of  $y_1$  that is extreme in one direction is, on average, more likely to be paired with a value of  $y_2$  that is less extreme in that direction, and so the averages will have less extreme values. Finally, you will notice that the sampling distribution also looks normal. In this particular case the sampling distribution is actually normal but for other problems it might not be so, although often it's close enough that it's safe to treat it as normal.

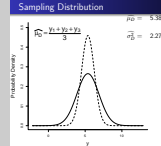


# Sampling Distribution



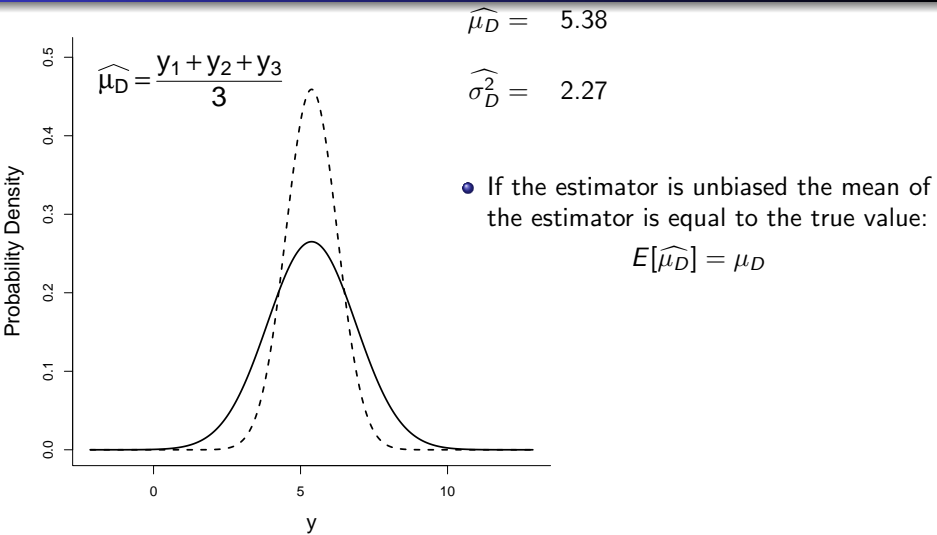
## Lecture 1

### Sampling Distribution



If we had three observations then our sampling distribution gets even narrower. Again, this is what we would like to see; as we collect more data the variability in our estimate should go down, and with very large sample sizes the variance becomes so small that from a practical perspective the estimate is equal to the true value.

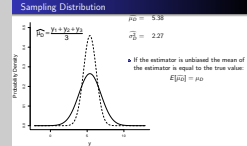
# Sampling Distribution



## Lecture 1

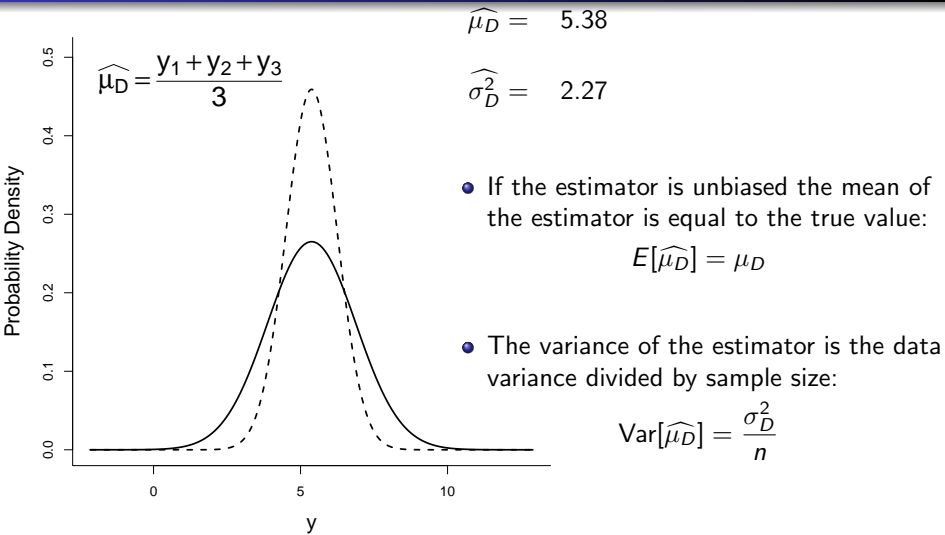
### Sampling Distribution

For the mean of a normal distribution ( $\mu_D$ ) the ML estimate ( $\hat{\mu}_D$ ) is the mean of the data  $\bar{y}$ . This estimator is unbiased because the expected value of the estimate  $E[\hat{\mu}_D]$  is equal to the true value. The mean of the dashed distribution (the sampling distribution) is equal to the mean of the true data distribution (the black line - don't forget at this point we are assuming our ML estimate of the distribution is the true distribution).





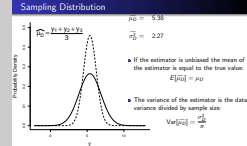
# Sampling Distribution



## Lecture 1

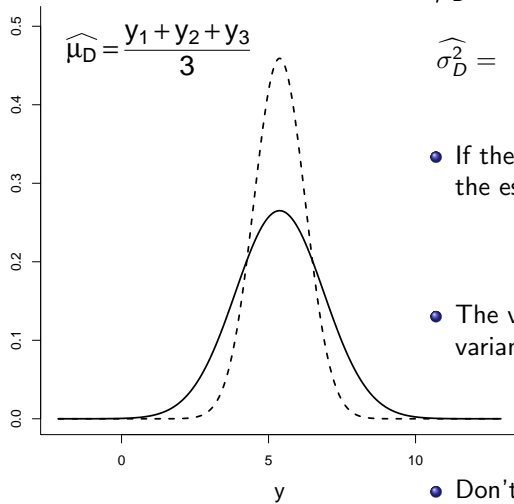
### Sampling Distribution

The variance of the estimator - the sampling variance - is equal to the true underlying variance of the data distribution divided by the number of observations  $n$  used to obtain the estimate. The problem of course is that we don't know the true underlying variance and so what we have to do



# Sampling Distribution

Probability Density



$$\hat{\mu}_D = 5.38$$

$$\hat{\sigma}_D^2 = 2.27$$

- If the estimator is unbiased the mean of the estimator is equal to the true value:

$$E[\hat{\mu}_D] = \mu_D$$

- The variance of the estimator is the data variance divided by sample size:

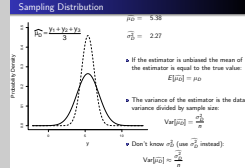
$$\text{Var}[\hat{\mu}_D] = \frac{\sigma_D^2}{n}$$

- Don't know  $\sigma_D^2$  (use  $\hat{\sigma}_D^2$  instead):

$$\text{Var}[\hat{\mu}_D] \approx \frac{\hat{\sigma}_D^2}{n}$$

## Lecture 1

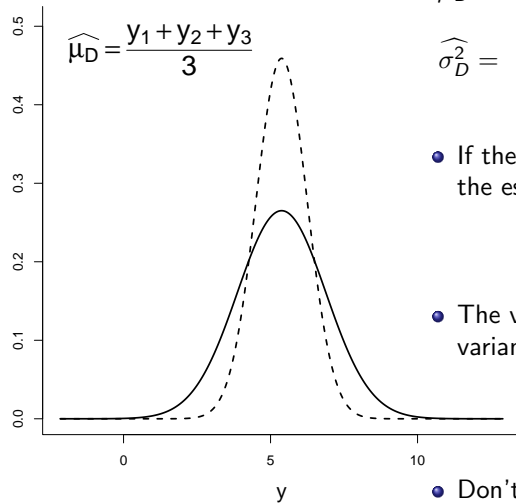
### Sampling Distribution



is replace it with an estimate - in this case our maximum likelihood estimate - to get an approximate sampling variance.

# Sampling Distribution

Probability Density



$$\hat{\mu}_D = 5.38$$

$$\hat{\sigma}_D^2 = 2.27$$

- If the estimator is unbiased the mean of the estimator is equal to the true value:

$$E[\hat{\mu}_D] = \mu_D$$

- The variance of the estimator is the data variance divided by sample size:

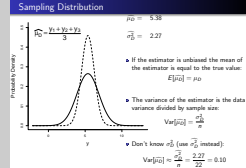
$$\text{Var}[\hat{\mu}_D] = \frac{\sigma_D^2}{n}$$

- Don't know  $\sigma_D^2$  (use  $\hat{\sigma}_D^2$  instead):

$$\text{Var}[\hat{\mu}_D] \approx \frac{\hat{\sigma}_D^2}{n} = \frac{2.27}{22} = 0.10$$

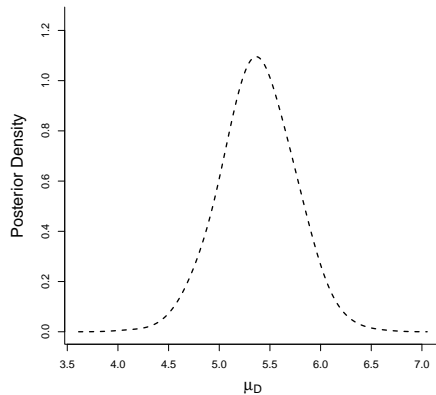
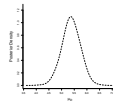
## Lecture 1

### Sampling Distribution



So the sampling variance for the true underlying mean score of grumpy photos is approximately our estimate of the variance (2.27) divided through by 22: 0.10. So the sampling variation is reasonably small.

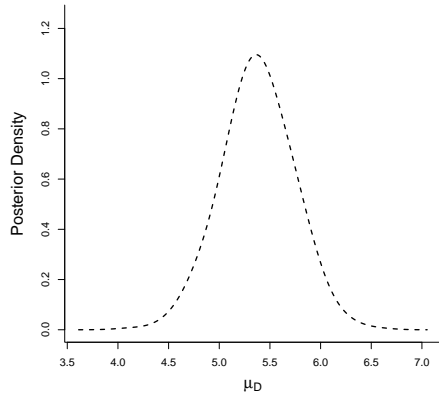
### Sampling Distribution versus Posterior Distribution



This is the posterior distribution for the mean of the distribution describing the scores of grumpy photos. We saw it earlier - and I noted that it has a very easy interpretation; it describes our epistemic uncertainty about the true value.

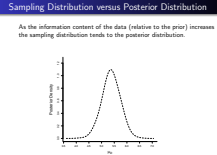
# Sampling Distribution versus Posterior Distribution

As the information content of the data (relative to the prior) increases the sampling distribution tends to the posterior distribution.



## Lecture 1

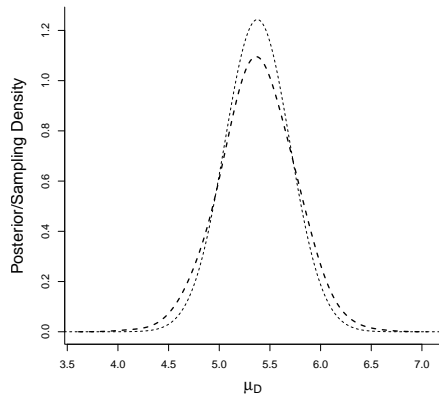
### Sampling Distribution versus Posterior Distribution



As the amount of data increases the sampling distribution converges on the posterior distribution, and even with a modest sample size we can see

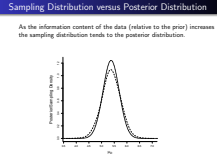
# Sampling Distribution versus Posterior Distribution

As the information content of the data (relative to the prior) increases the sampling distribution tends to the posterior distribution.



## Lecture 1

### Sampling Distribution versus Posterior Distribution



that the two largely agree in this instance if we make the prior vague. And so I think that although sampling distributions and posterior distributions are fundamentally different things I think it is OK to think about sampling distributions in this straightforward way.

# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)

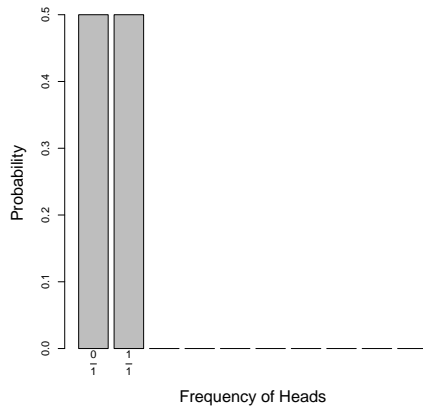
## Lecture 1

### └ Sampling Distribution tends to a Normal

Another property of sampling distributions is that as the amount of information in the data about a parameter increases they tend to normal distributions. In some cases they may do so quite slowly, but often they can be safely treated as such. It's also important to realise that this will happen even if the data aren't normally distributed. For example, lets imagine we were trying to estimate the probability of flipping a head from this coin (rather than assuming it's 0.5). If we flipped it once the outcome could either be a tail in which case our best estimate of the probability would be 0 or it could be a head in which case our best estimate of the probability would be 1.

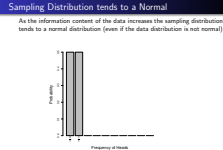
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



## Lecture 1

└ Sampling Distribution tends to a Normal

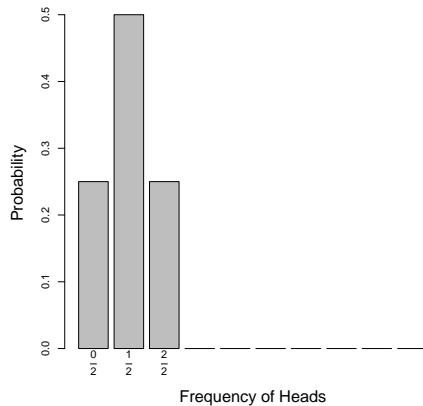


If the coin was in reality fair, then the sampling distribution would look like this. Half the time we would estimate 0 and half the time we would estimate 1. The estimator is unbiased - the average of the estimates would be 0.5, but the distribution is far from being Gaussian - it just has two point masses (i.e. two values where the probability is non-zero).



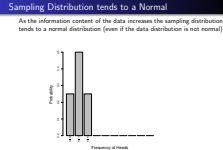
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



## Lecture 1

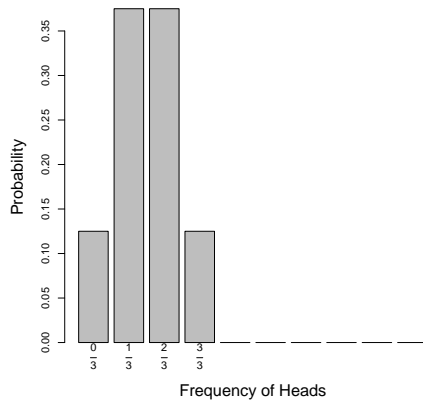
Sampling Distribution tends to a Normal



If we flipped the coin twice we have three possible outcomes (0,1, or 2 heads) and therefore three possible estimates (0, 1/2 or 1).

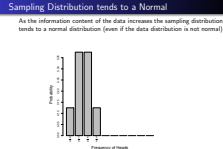
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



## Lecture 1

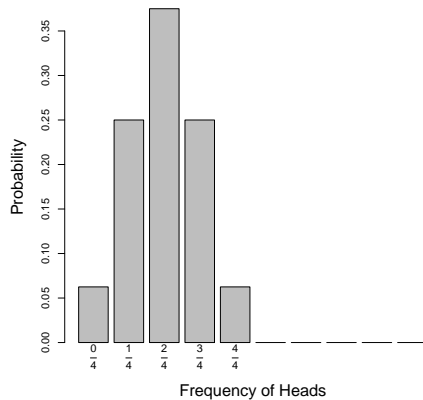
Sampling Distribution tends to a Normal



With three flips we have four possible estimates (0, 1/3, 2/3 or 1).

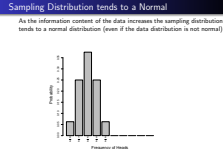
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



## Lecture 1

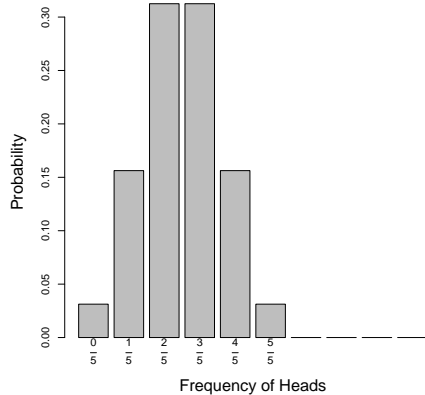
└ Sampling Distribution tends to a Normal



and so on

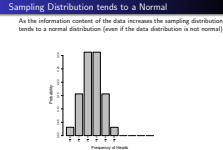
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



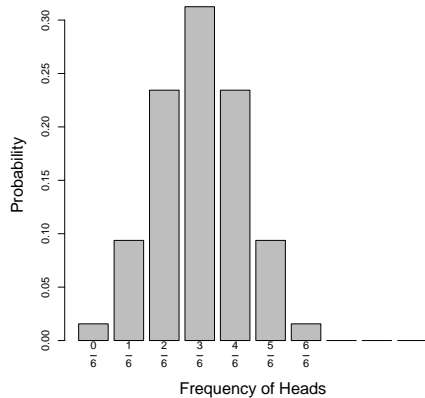
## Lecture 1

Sampling Distribution tends to a Normal



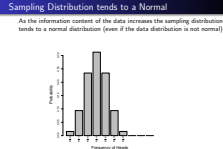
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



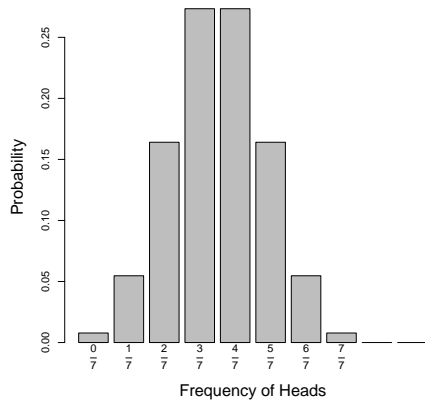
## Lecture 1

Sampling Distribution tends to a Normal



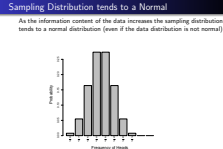
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



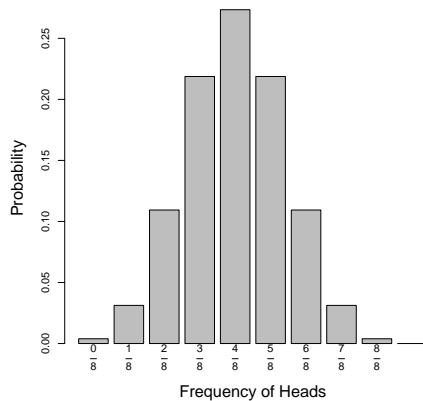
## Lecture 1

Sampling Distribution tends to a Normal



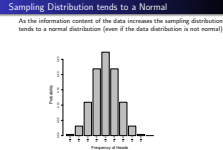
# Sampling Distribution tends to a Normal

As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



## Lecture 1

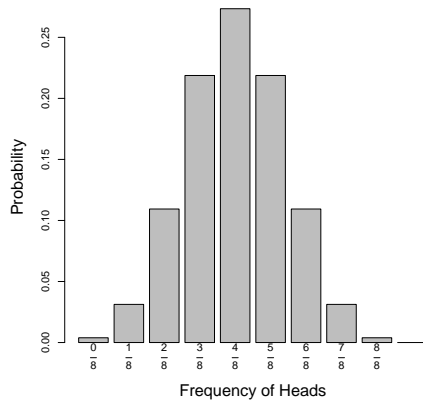
Sampling Distribution tends to a Normal



And you can see that even when our data are far from normal (for example, either zero's or ones) the sampling distribution of an underlying parameter (the probability of flipping a head) starts to look normal pretty rapidly

# Sampling Distribution tends to a Normal

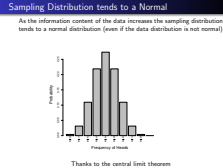
As the information content of the data increases the sampling distribution tends to a normal distribution (even if the data distribution is not normal)



Thanks to the central limit theorem

## Lecture 1

Sampling Distribution tends to a Normal



and this is thanks to the central limit theorem; once you take the average of many numbers the distribution of those averages tend to a normal distribution even if the underlying numbers have a wacky distribution, such as zero or one as here,



### └─ Uncertainty and Distributions: mini-quiz

We've covered quite a bit of ground so far, and it has been fairly abstract, so lets just recap on the key points we've learnt.

- Uncertainty

### └─ Uncertainty and Distributions: mini-quiz

The first thing that we had to get straight in our heads is that when we use the word uncertainty or probability we're actually talking about two separate things and it's easy to get confused. We have aleatoric uncertainty which is the type of uncertainty that comes with rolling a dice or tossing a coin, and there's epistemic uncertainty which is to do with knowledge.

Now, birdwatchers tend not to like cats much (or dogs)

# Uncertainty and Distributions: mini-quiz

- Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

## Lecture 1

### └─ Uncertainty and Distributions: mini-quiz

and I'm sorry to tell you that a particularly rabid birdwatcher has been chopping the heads off half the cats they meet.

- Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

what is the probability a cat they meet will die?

what is the probability the cat they met is dead?

### └─ Uncertainty and Distributions: mini-quiz

Here are two questions. What type of uncertainty do you think these questions are concerned with?

Q1 is aleatoric and Q2 is epistemic; it is the same as the coin example I gave earlier. Q1 is referring to something that might happen hypothetically and we know that the probability of it being actual is 0.5. Q2 is referring to something that has happened (the cat is either dead or alive) and the uncertainty is because you haven't seen the headless (or headed) cat.

- Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

what is the probability a cat they meet will die?

what is the probability the cat they met is dead?

# Uncertainty and Distributions: mini-quiz

- Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

what is the probability a cat they meet will die?

what is the probability the cat they met is dead?

- Distributions

## Lecture 1

### └─ Uncertainty and Distributions: mini-quiz

The other thing we covered is distributions, and in particular three types of distribution. The data distribution, which is the distribution from which our observations arose. The posterior distribution which characterises our epistemic uncertainty about something (often a parameter) and the sampling distribution. The sampling distribution is the hardest; it is the distribution of our estimates had we repeated our experiment a very large number of times.

- Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

what is the probability a cat they meet will die?

what is the probability the cat they met is dead?

- Distributions

# Uncertainty and Distributions: mini-quiz

- Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

what is the probability a cat they meet will die?

what is the probability the cat they met is dead?

- Distributions

They've given 10 cats polonium, 8 are dead.

## Lecture 1

### └─ Uncertainty and Distributions: mini-quiz

Now the head chopping got boring so the birdwatcher also gave 10 cats polonium to see what would happen. 8 died.

- Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

what is the probability a cat they meet will die?

what is the probability the cat they met is dead?

- Distributions

They've given 10 cats polonium, 8 are dead.

# Uncertainty and Distributions: mini-quiz

## • Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

what is the probability a cat they meet will die?

what is the probability the cat they met is dead?

## • Distributions

They've given 10 cats polonium, 8 are dead.

what is the probability ( $p$ ) a cat will die after polonium?

The home office has let them redo the experiment a gazillion times  
In how many experiments did 3 out of 10 cats die?

They have 10 trials and 'success' with probability  $p$ .

## Lecture 1

### └─ Uncertainty and Distributions: mini-quiz

Uncertainty and Distributions: mini-quiz

- Uncertainty
  - A rabid birdwatcher is chopping the heads off half the cats they meet.
    - what is the probability a cat they meet will die?
    - what is the probability the cat they met is dead?
- Distributions
  - They've given 10 cats polonium, 8 are dead.
    - what is the probability ( $p$ ) a cat will die after polonium?
  - The home office has let them redo the experiment a gazillion times
    - In how many experiments did 3 out of 10 cats die?
  - They have 10 trials and 'success' with probability  $p$ .

Here are three more questions/statements. What type of distribution do you think I'm talking about?

## • Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

*aleatoric*: what is the probability a cat they meet will die?

*epistemic*: what is the probability the cat they met is dead?

## • Distributions

They've given 10 cats polonium, 8 are dead.

what is the probability ( $p$ ) a cat will die after polonium?

The home office has let them redo the experiment a gazillion times  
In how many experiments did 3 out of 10 cats die?

They have 10 trials and 'success' with probability  $p$ .

## └─ Uncertainty and Distributions: mini-quiz

Uncertainty and Distributions: mini-quiz

- **Uncertainty**  
A rabid birdwatcher is chopping the heads off half the cats they meet.  
*aleatoric*: what is the probability a cat they meet will die?  
*epistemic*: what is the probability the cat they met is dead?
- **Distributions**  
They've given 10 cats polonium, 8 are dead.  
what is the probability ( $p$ ) a cat will die after polonium?  
The home office has let them redo the experiment a gazillion times  
In how many experiments did 3 out of 10 cats die?  
They have 10 trials and 'success' with probability  $p$ .

Uncertainty Questions:

Q1 is aleatoric and Q2 is epistemic; it is the same as the coin example I gave earlier. Q1 is referring to something that might happen hypothetically and we know that the probability of it being actual is 0.5. Q2 is referring to something that has happened (the cat is either dead or alive) and the uncertainty is because you haven't seen the headless (or headed) cat.



## • Uncertainty

A rabid birdwatcher is chopping the heads off half the cats they meet.

*aleatoric*: what is the probability a cat they meet will die?

*epistemic*: what is the probability the cat they met is dead?

## • Distributions

They've given 10 cats polonium, 8 are dead.

*posterior*: what is the probability ( $p$ ) a cat will die after polonium?

*sampling*: The home office has let them redo the experiment a gazillion times  
In how many experiments did 3 out of 10 cats die?

*data*: They have 10 trials and 'success' with probability  $p$ .

## └─ Uncertainty and Distributions: mini-quiz

Uncertainty and Distributions: mini-quiz

- **Uncertainty**  
A rabid birdwatcher is chopping the heads off half the cats they meet.  
*aleatoric*: what is the probability a cat they meet will die?  
*epistemic*: what is the probability the cat they met is dead?
- **Distributions**  
They've given 10 cats polonium, 8 are dead.  
*posterior*: what is the probability ( $p$ ) a cat will die after polonium?  
*sampling*: The home office has let them redo the experiment a gazillion times  
In how many experiments did 3 out of 10 cats die?  
*data*: They have 10 trials and 'success' with probability  $p$ .

Distribution Questions: Q1 posterior distribution, Q2 sampling distribution and Q3 data distribution.

Q1 is about the posterior because we're talking directly about the (epistemic) uncertainty of an unknown parameter. Q2 is a sampling distribution as it describes the distribution of the estimates of the probability in hypothetical reruns of the experiment. In experiments where 3/10 cats die, the best (maximum likelihood) estimate ( $\hat{p}$ ) is 0.3 and the probability of this happening is equal to its frequency when rerunning the experiment. Statement 3 describes the binomial distribution from which our data arise.

$$\widehat{\mu}_D = 5.376 \quad \widehat{\sigma}_D^2 = 2.266 \quad \text{Var}[\widehat{\mu}_D] = 0.103$$

└ 1m

OK - so we've obtained our maximum likelihood estimators: our best estimate of the mean grumpiness score for grumpy photos is 5.4 and our best estimate of the variance in grumpiness score for grumpy photos is around 2.3. And we estimated the sampling variance of our estimate (for the mean) as 0.10. It's often easier to think about standard deviations rather than variances

$$\widehat{\mu}_D = 5.376 \quad \widehat{\sigma}_D^2 = 2.266 \quad \text{Var}[\widehat{\mu}_D] = 0.103$$

$$\widehat{\sigma}_D = 1.505 \quad \text{SD}[\widehat{\mu}_D] = 0.321$$

└ 1m

so we can just take the square root of these quantities. If the data distribution was normal and these estimates were the actual values then approximately 70% of data points are going to lie within one standard deviation of the mean ( $5.38 \pm 1.51$ ; so somewhere between 3.87 and 6.88) and approximately 95% are going to lie within two standard deviations of the mean; so somewhere between  $5.38 \pm 2 \times 1.51$  (so somewhere between 2.37 and 8.39). Likewise, our best estimate of the mean of our data distribution is the maximum likelihood estimator (5.38) and it is unlikely that the true underlying mean is going to lie further than two sampling standard deviations from this number ( $5.38 \pm 2 \times 0.32$ ; so somewhere between 4.73 and 6.02).

I didn't show you how we found the parameters of the data distribution that maximise the probability of seeing the data (the maximum likelihood estimate) and I don't think it is necessary for you to understand optimisation algorithms.

$$\widehat{\mu}_D = 5.376 \quad \widehat{\sigma}_D^2 = 2.266 \quad \text{Var}[\widehat{\mu}_D] = 0.103$$

$$\widehat{\sigma}_D = 1.505 \quad \text{SD}[\widehat{\mu}_D] = 0.321$$

```
> photo_m1 <- lm(y ~ 1, data = subset(photo_long,
+   type == "grumpy"))
```

└─lm

However, fitting a linear model to the data finds the maximum likelihood estimates under the assumption that the data distribution is normal (also known as Gaussian). So people should be familiar with this bit of R-code. The function `lm` fits a linear model. The first argument is a formula that specifies the model. We start by specifying the response variable (in our case `y`) followed by a tilde (`~`). The tilde is often used as shorthand for '*is distributed as*'. On the right-hand side of the formula we specify how we would like the mean of our normal distribution to change. We'll cover this in more detail later, but in this model I've simply put a one. In R a 1 in a model formula stands for the intercept. It is the mean value of the response variable you would expect if all other predictor variables were set to zero. In this model we don't have any other predictor variables so the intercept is simply the mean response. We've also sub-setted our data so we are only analysing those photos that were made under grumpy conditions.

```
lm
μ̂_D = 5.376  σ̂_D^2 = 2.266  Var[μ̂_D] = 0.103
σ̂_D = 1.505  SD[μ̂_D] = 0.321
> photo_m1 <- lm(y ~ 1, data = subset(photo_long,
+   type == "grumpy"))
```

$$\widehat{\mu}_D = 5.376 \quad \widehat{\sigma}_D^2 = 2.266 \quad \text{Var}[\widehat{\mu}_D] = 0.103$$

$$\widehat{\sigma}_D = 1.505 \quad \text{SD}[\widehat{\mu}_D] = 0.321$$

```
> photo_m1 <- lm(y ~ 1, data = subset(photo_long,
+   type == "grumpy"))
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.3759	0.3209	16.75	1.25e-13 ***

---

Signif. codes:

0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.505 on 21 degrees of freedom

└─lm

```
lm
      mu_D = 5.376   sigma_D^2 = 2.266   Var[mu_D] = 0.103
      sigma_D = 1.505   SD[mu_D] = 0.321
> photo_m1 <- lm(y ~ 1, data = subset(photo_long,
+   type == "grumpy"))
> summary(photo_m1)
Residuals:
    Min       1Q   Median       3Q      Max
-3.196 -1.173 -0.089  1.105  2.682
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.3759    0.3209   16.75 1.25e-13 ***
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.505 on 21 degrees of freedom
```

If you are not used to the output it's a little bewildering; and whether you are used to it not, there's these little stars to drag your attention away from the more important stuff. You can see that the estimate of the intercept coincides with our maximum likelihood estimate of it, and our estimate of the standard deviation is equal to what is called the residual standard error. It's a shame they call it a standard error; it is the standard deviation of the residuals (the standard deviation of the data around their expected values). Finally we have our sampling standard deviation of our mean estimate, which is usually referred to as the standard error.

There's also a few other bits of information. We have a rough summary of how the residuals are distributed and we have a t-value followed by a p-value. Lets focus on the residuals first.

# 1m: Residuals

```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682

## Lecture 1

└ 1m: Residuals

This is some basic summary information about the distribution of the residuals; the minimum, median and maximum values which you will be familiar with, but also the first and third quartiles.

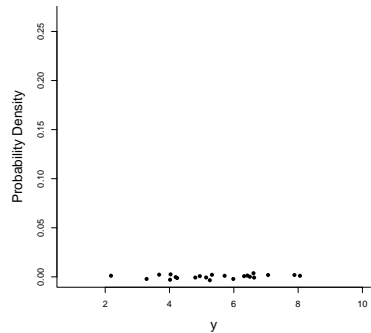
```
1m: Residuals
> summary(photo_m1)
Residuals:
   Min     1Q   Median     3Q      Max
-3.196 -1.173 -0.089  1.105  2.682
```

# 1m: Residuals

```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682

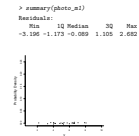


## Lecture 1

└─ 1m: Residuals

These are our data

1m: Residuals

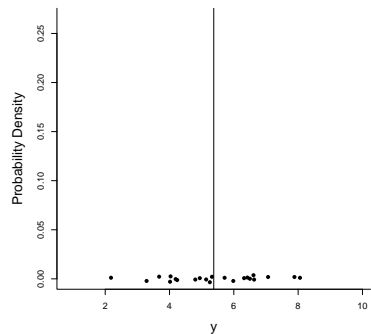


# 1m: Residuals

```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



## Lecture 1

└─ 1m: Residuals

1m: Residuals

```
> summary(photo_m1)
Residuals:
    Min       1Q   Median       3Q      Max
-3.196 -1.173 -0.089  1.105  2.682
```



And this is our maximum likelihood estimate of the mean. The residuals are deviations from this line and will have a mean of *exactly* zero by construction.

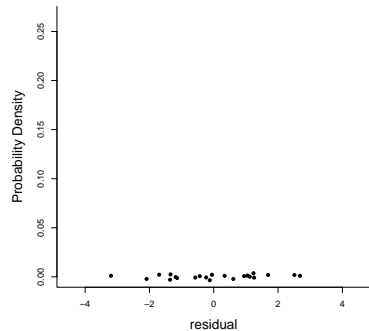


# 1m: Residuals

```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682

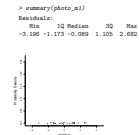


## Lecture 1

└─ 1m: Residuals

Here we have the deviations - the residuals.

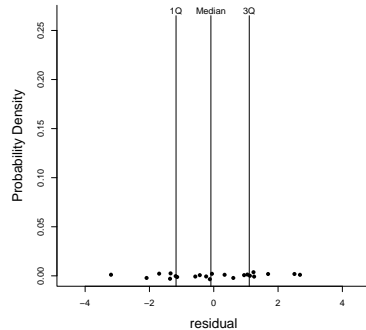
1m: Residuals



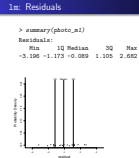
```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



### 1m: Residuals

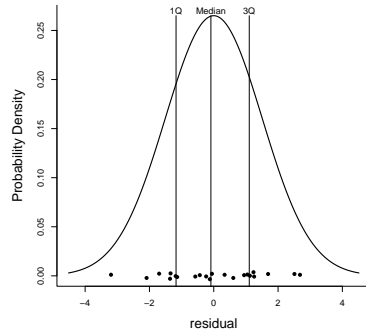


If we sort our residuals, the first quartile is the point at which we have seen 25% of them, the median is the point at which we have seen half the residuals (the second quartile) and the third quartile is the point at which we have seen 75% of the residuals. Because we've assumed the residuals are normal, and the normal is symmetric, we would want the median to be close to zero and we'd like the quartiles to be roughly the same but opposite in sign. It would be nice if this was also true of the minima and maxima but these are often quite noisy and even quite different values might still be consistent with a normal distribution. This summary of the residual distribution is not very useful I think; I'd be amazed if many people even bothered to look at it.

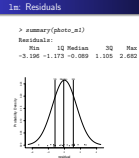
```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



### 1m: Residuals

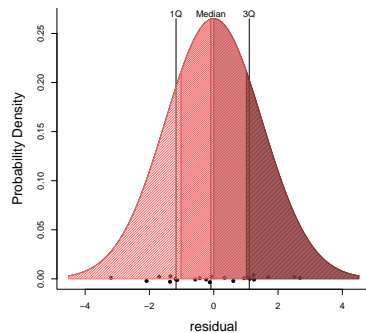


It's much easier to assess graphically whether the residuals look close to being normal or not. This is our maximum likelihood estimate for the distribution of the residuals; it has a mean of zero and a variance equal to our estimate 2.27. What we can do is say if our data really were generated from this distribution where do we *expect* the 1st, 2nd and 3rd quartiles to be? As I mentioned before the area under this curves equals 1, and so what we could do is chop it up into four regions of equal volume, within each of which we expect a quarter of our residuals to lie.

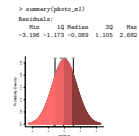
```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



└─ 1m: Residuals

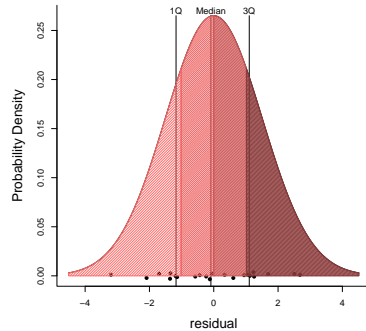


If we do that for our model it looks pretty good: the three places where we expect to chop the distribution in four are pretty close to the empirical quartiles.

```
> summary(photo_m1)
```

Residuals:

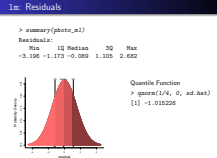
Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



Quantile Function

```
> qnorm(1/4, 0, sd.hat)
[1] -1.015226
```

1m: Residuals

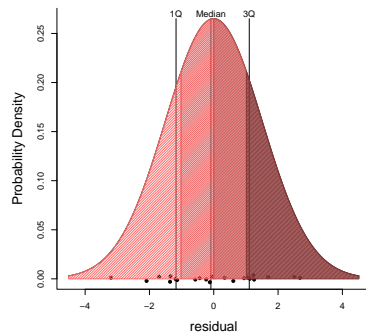


If we want to find the number below which we should have seen x% of the data we use the quantile function. So to get the first quartile I specify a probability of a quarter and it returns the value below which 25% of the data should lie (here `sd.hat` is the estimate of the residual standard deviation - 1.505).

```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



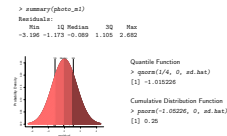
Quantile Function

```
> qnorm(1/4, 0, sd.hat)
[1] -1.015226
```

Cumulative Distribution Function

```
> pnorm(-1.05226, 0, sd.hat)
[1] 0.25
```

## 1m: Residuals

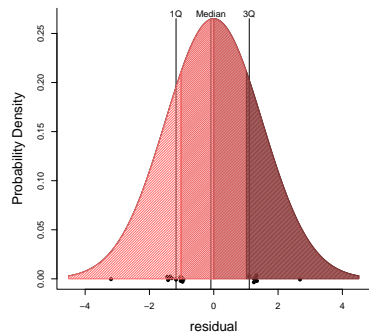


The inverse of the quantile function is called the cumulative density function (or cumulative mass function if the outcome is discrete) and here we give it some value and it tells us what is the probability of an observation being smaller than this. These are super useful functions. Now assessing whether the residuals are consistent with a normal by just looking at these three numbers is a pretty blunt diagnostic. For example,

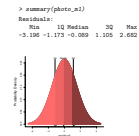
```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



└ 1m: Residuals

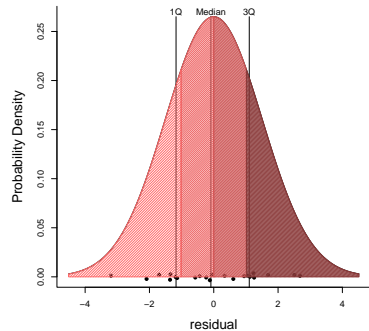


These residuals have the same quartiles and range as our original residuals but clearly they do not conform very well to the normal.

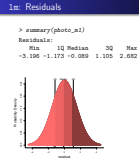
```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682



└─ 1m: Residuals



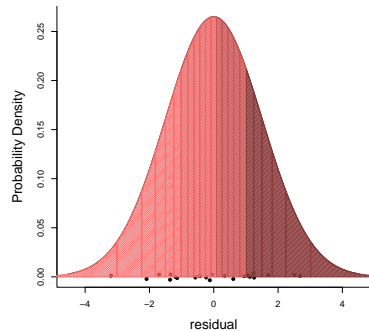
If we take our original data, what we can do



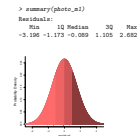
```
> summary(photo_m1)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.196	-1.173	-0.089	1.105	2.682

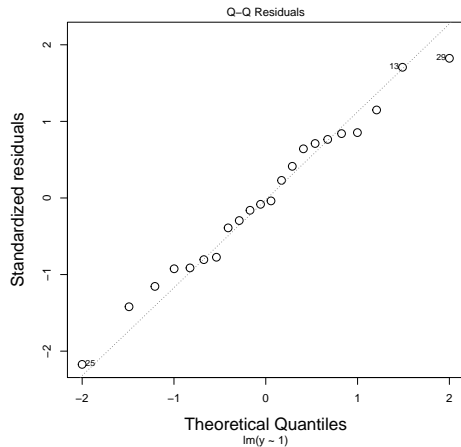


└ 1m: Residuals

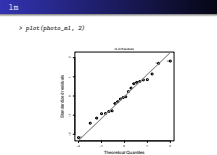


is divide the distribution up into as many data points we have (in this case 22) and split our estimated distribution up into 22 equally sized areas and see if the expected quantiles coincide with the residuals.

```
> plot(photo_m1, 2)
```



```
lm
```



This is called a quantile plot or a qq-plot. On the x-axis we have the theoretical quantiles; the places where we think we should cut the distribution into 22 pieces and on the y-axis we have the actual residuals from our model<sup>[1]</sup>. We can see that the points are lying pretty much on the 1:1 line<sup>[2]</sup> indicating that what we expect from the normal distribution is pretty much what we see. There's a little bit of deviation at the extremes, but that is normal. Extreme values tend to be quite noisy.

<sup>[1]</sup> Note that the residuals have been standardized (divided by the residual standard deviation in this instance) and so the theoretical distribution is the unit normal (mean of zero and standard deviation of 1). This is because  $\text{Var}(cx)$  where  $c$  is a constant is equal to  $c^2\text{Var}(x)$ . If  $c = 1/\sigma$  (i.e. we divide  $x$  by its standard deviation) we get  $\text{Var}(x/\sigma) = \text{Var}(x)/\sigma^2 = \sigma^2/\sigma^2 = 1$ .

<sup>[2]</sup> Note that the dashed diagonal line is NOT the 1:1 line, but a line going through the pair of points at the first and third quantile. This can be very misleading in some models we'll cover later (e.g. overdispersion in a Poisson model) and it's not clear to me why the 1:1 line is not plotted.

# 1m: Confidence Intervals

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

## Lecture 1

### └─ 1m: Confidence Intervals

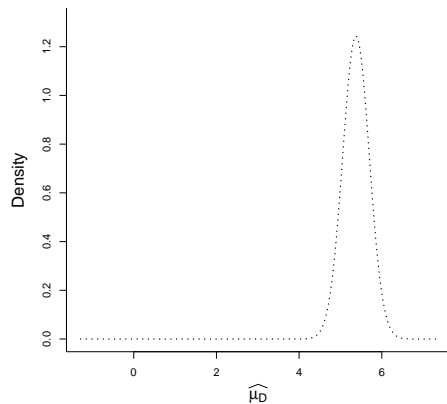
After a summary of the residuals we get the coefficient table. We've got our maximum likelihood estimate of the mean, we've got the standard deviation of the sampling distribution (the standard error) and then we've got a t-value and some odd symbol ( $\Pr(>|t|)$ ) that designates something you're probably not too sure about but you do know this is the beloved p-value.

```
1m: Confidence Intervals
> coef(summary(photo_m1))
              Estimate Std. Error t value    Pr(>|t|)
(Intercept)  5.375865   0.3209046  16.75222 1.254999e-13
```

# 1m: Confidence Intervals

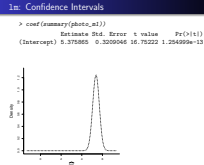
```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



## Lecture 1

### 1m: Confidence Intervals

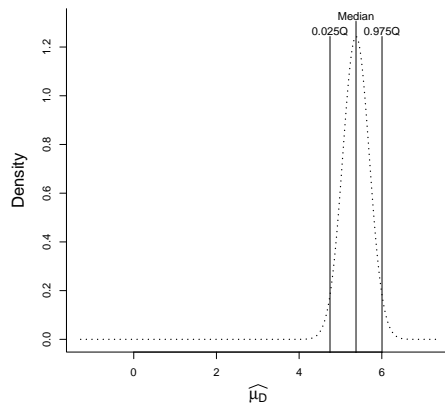


We can plot the sampling distribution around our estimate since we know the mean (the estimate) and the standard deviation (the standard error) and we also saw earlier that the sampling distribution of our mean will be normal.

# 1m: Confidence Intervals

```
> coef(summary(photo_m1))
```

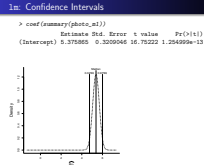
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



## Lecture 1

### 1m: Confidence Intervals

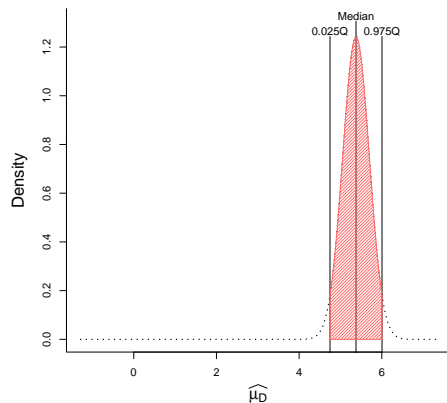
We could also find the 2.5% and 97.5% quantiles of our sampling distribution.



# 1m: Confidence Intervals

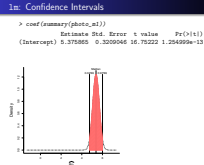
```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



## Lecture 1

### 1m: Confidence Intervals

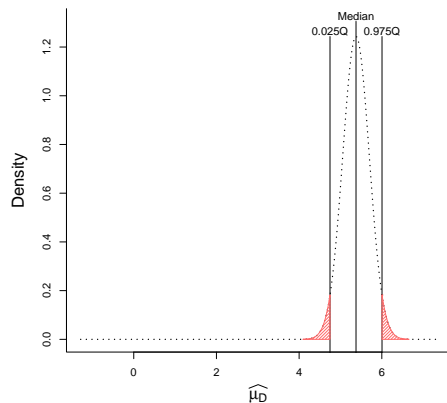


These are the values in which we expect 95% of the estimates to lie had we repeated the experiment and if our estimate was the true value. We've also made sure that the two tails contain the same amount of probability. Conversely,

# 1m: Confidence Intervals

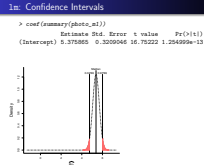
```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



## Lecture 1

### 1m: Confidence Intervals

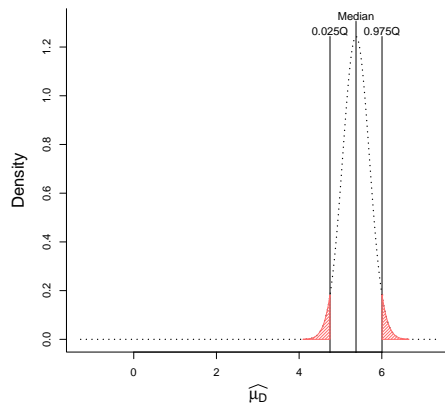


We expect 5% of the estimates to lie outside this region, with half of them in the lower tail and half of them in the upper tail. This interval you will be familiar with, it's the 95% confidence interval

# 1m: Confidence Intervals

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



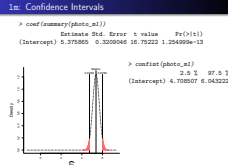
```
> confint(photo_m1)
```

	2.5 %	97.5 %
(Intercept)	4.708507	6.043222

## Lecture 1

### 1m: Confidence Intervals

and we can simply use the R function `confint` to obtain it. We can also do it 'by hand'

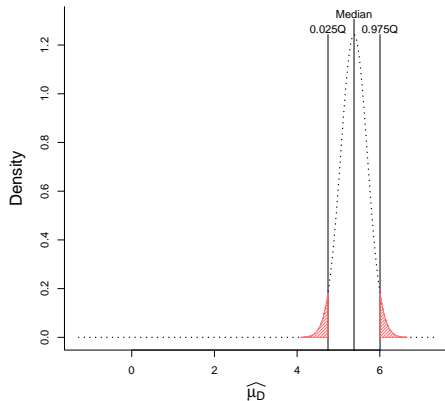




# 1m: Confidence Intervals

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



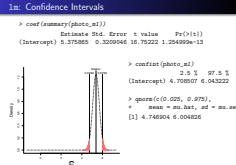
```
> confint(photo_m1)
```

	2.5 %	97.5 %
(Intercept)	4.708507	6.043222

```
> qnorm(c(0.025, 0.975),  
+       mean = mu.hat, sd = mu.se)  
[1] 4.746904 6.004826
```

## Lecture 1

### 1m: Confidence Intervals

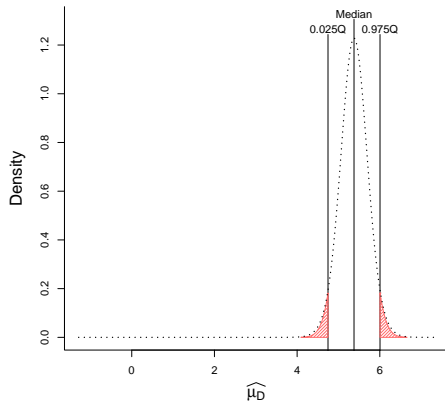


by using the quantile function we used before: so we can work out the value under which the lower 2.5% of estimates should lie, and the value under which the lower 97.5% of estimates should lie. And .... Oh dear .... close but not the same. The interval we have worked out by hand is slightly tighter. The reason for this is that I lied to you. I told you that the sampling distribution of the mean of a normal distribution was itself normal. However, that is only true if we know the variance of the data distribution. But we don't know it: when we were calculating the standard error I replaced the true data variance  $\sigma_D^2$  with its estimate  $\hat{\sigma}_D^2$  and told you not to worry about it. Now if by chance our estimate of the data variance was too small, then the variance of the sampling distribution (the width of the pictured distribution) should really have been larger, and if by chance our estimate of the data variance was too large, then the variance of the sampling distribution is actually a little bit smaller than we've pictured here. We could imagine having a series of normal sampling distributions all with the same mean but a range of plausible standard deviations.

# 1m: Confidence Intervals

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



```
> confint(photo_m1)
```

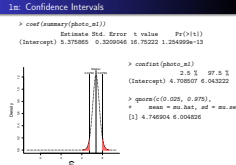
	2.5 %	97.5 %
(Intercept)	4.708507	6.043222

```
> qnorm(c(0.025, 0.975),  
+       mean = mu.hat, sd = mu.se)
```

```
[1] 4.746904 6.004826
```

## Lecture 1

### 1m: Confidence Intervals



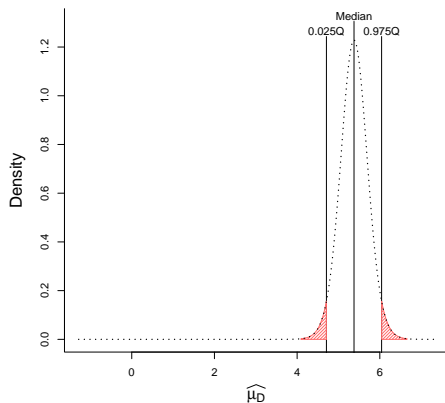
If we blended these different plausible sampling distributions (a blended distribution is called a compound or mixture distribution) then we would end up with a new sampling distribution represented by this new dotted line. You probably didn't notice the difference but if you look carefully you will see that our new distribution has slightly higher density at extreme values - it has fatter tails (the red hatched area is consistently below the new density function). This new blended distribution is called a t-distribution.<sup>[1]</sup>

<sup>[1]</sup> Actually this is a mean-shifted and scaled t-distribution. The standard t-distribution has only one parameter - the degrees of freedom, often denoted  $\nu$ . It has a mean of 0 and standard deviation of  $\sqrt{\nu/(\nu-2)}$ . Mean-shifting just means we shift the distribution to a new mean (specified by `mean` in the `t.scaled` family of functions available in the library `metRology`) and we scale it by a parameter  $s$  such that the new standard deviation is  $s\sqrt{\nu/(\nu-2)}$  ( $s$  is specified as `sd` in the `t.scaled` family of functions, but this invites confusion because this is not the standard deviation of the distribution). An alternative way to think about it is if  $X$  is a t-distributed random variable, then  $sX + \text{mean}$  has a mean-shifted and scaled t-distribution.

# 1m: Confidence Intervals

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



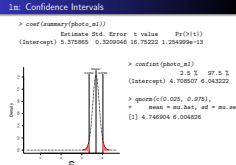
```
> confint(photo_m1)
```

	2.5 %	97.5 %
(Intercept)	4.708507	6.043222

```
> qnorm(c(0.025, 0.975),  
+       mean = mu.hat, sd = mu.se)  
[1] 4.746904 6.004826
```

## Lecture 1

### 1m: Confidence Intervals

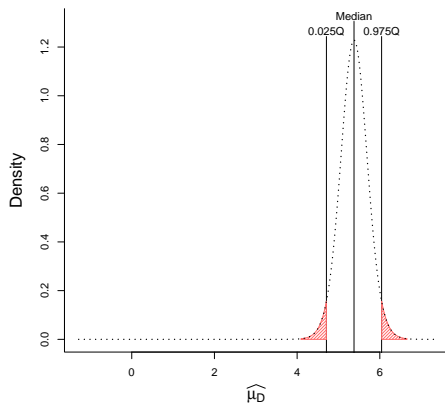


What we can do is recalculate the 2.5% and 97.5% quantiles for this new distribution, so work out where these lines need to be so each tail contains 2.5% of the probability. Again, you probably didn't even notice the slight shift in the quantiles to more extreme values. We can also work it out 'by hand' so we can feel confident that we understand what the confint function is telling us.

# 1m: Confidence Intervals

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



```
> confint(photo_m1)
```

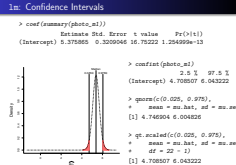
	2.5 %	97.5 %
(Intercept)	4.708507	6.043222

```
> qnorm(c(0.025, 0.975),  
+       mean = mu.hat, sd = mu.se)  
[1] 4.746904 6.004826
```

```
> qt.scaled(c(0.025, 0.975),  
+          mean = mu.hat, sd = mu.se,  
+          df = 22 - 1)  
[1] 4.708507 6.043222
```

## Lecture 1

### 1m: Confidence Intervals



And it gives us the expected answer. The function takes the same arguments as before, the estimate and its standard error, but also requires us to specify the degrees of freedom. As you will see when we move on to mixed models, it's sometimes hard to know what the degrees of freedom are, but in a simple model like this it is simply the number of observations used to estimate the residual variance minus the number of location parameters in the model (in this case 1, the intercept).

# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

## Lecture 1

```
1m: Hypothesis Testing  
  
> coef(summary(photo_m1))  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept) 5.375865  0.3209046 16.75222 1.254999e-13
```

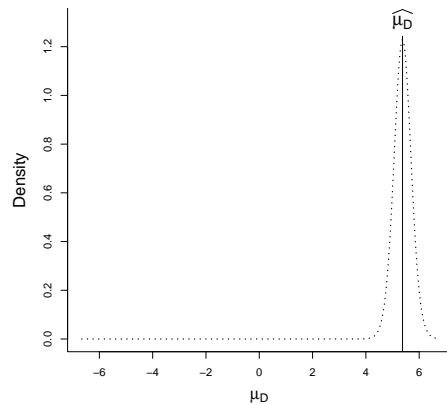
### └─ 1m: Hypothesis Testing

The next thing to understand is the t value and the p value.

# 1m: Hypothesis Testing

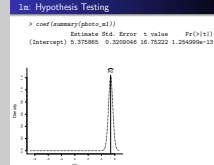
```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



## Lecture 1

### 1m: Hypothesis Testing

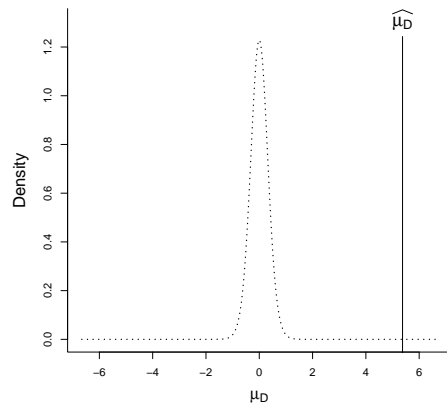


Here I've plotted our estimate of the mean and its sampling distribution, which as we saw is a scaled and mean-shifted t-distribution.

# 1m: Hypothesis Testing

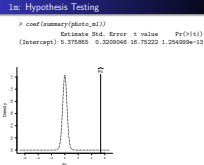
```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



## Lecture 1

### 1m: Hypothesis Testing

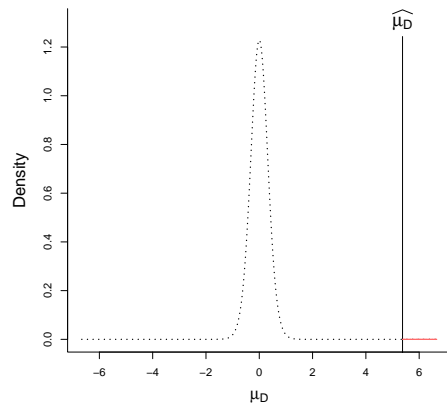


The estimate of the mean of a normal distribution has the nice property that the sampling distribution around the true mean would be the same no matter what the underlying true mean is. So for example if the true mean was zero the sampling distribution would look the same, just displaced. We could then ask, if zero really was the true mean (our null hypothesis) how likely is it that we would have obtained an estimate as far from zero as we did?

# 1m: Hypothesis Testing

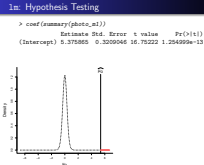
```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



## Lecture 1

└ 1m: Hypothesis Testing



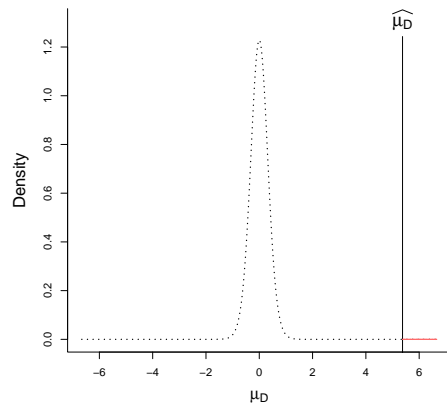
And the answer is we can just look to see how much probability there is in this tail. I've coloured the area in red, but because the probability is low you can barely see it.



# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

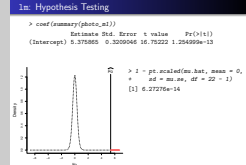
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13



```
> 1 - pt.scaled(mu.hat, mean = 0,  
+             sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```

## Lecture 1

### 1m: Hypothesis Testing



We can use the cumulative density function of the scaled t-distribution to evaluate this probability. we pass it our estimate and it tells us how likely we are to get a value *less* than this, so we need to one minus this number.

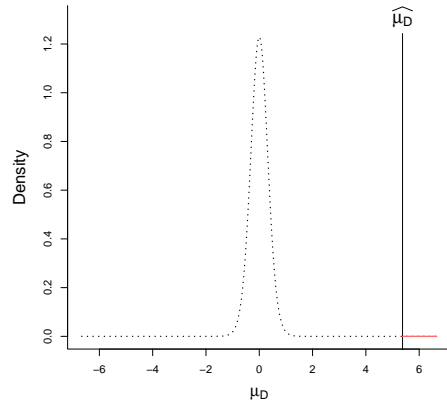
# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

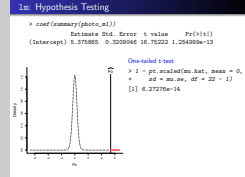
## One-tailed t-test

```
> 1 - pt.scaled(mu.hat, mean = 0,  
+             sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```



## Lecture 1

### 1m: Hypothesis Testing



This is what we call a one-tailed test. We have asked what is the probability of getting an estimate larger than this had the true value been zero.

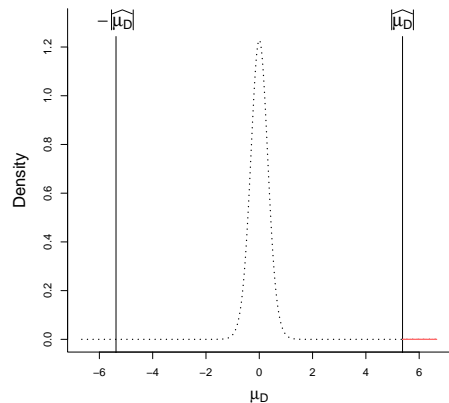
# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

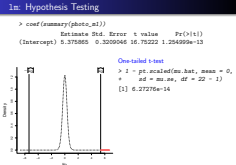
## One-tailed t-test

```
> 1 - pt.scaled(mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```



## Lecture 1

### 1m: Hypothesis Testing



However, what would have happened had we got the same estimate but opposite in sign?

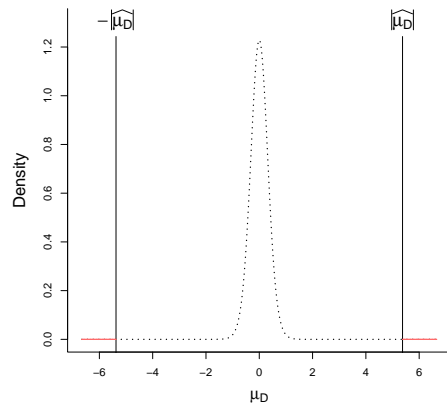
# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

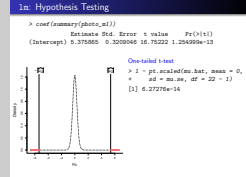
## One-tailed t-test

```
> 1 - pt.scaled(mu.hat, mean = 0,  
+             sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```



## Lecture 1

### 1m: Hypothesis Testing



The t-distribution is symmetric so the probability getting an estimate that is more negative than this, is the same as our previous probability - this area in red.

# 1m: Hypothesis Testing

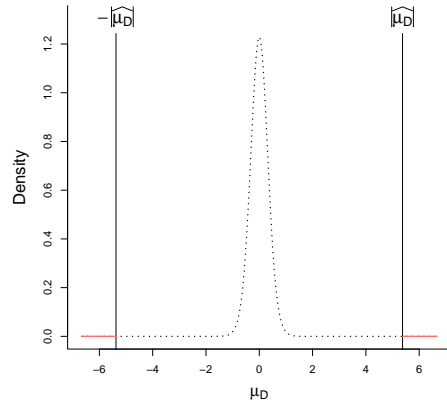
```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

## One-tailed t-test

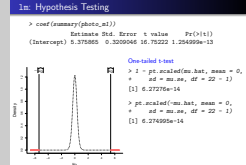
```
> 1 - pt.scaled(mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```

```
> pt.scaled(-mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.274995e-14
```



## Lecture 1

### 1m: Hypothesis Testing



And we can simply get this number by asking what is the probability of getting an estimate less than the negative of our estimate.

# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

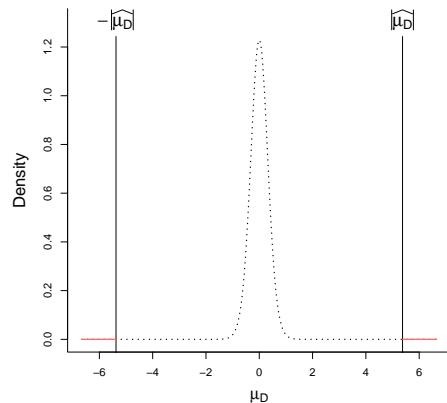
	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

## One-tailed t-test

```
> 1 - pt.scaled(mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```

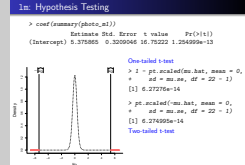
```
> pt.scaled(-mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.274995e-14
```

## Two-tailed t-test



## Lecture 1

### 1m: Hypothesis Testing



Most people don't want to rule out the possibility that deviations in either direction, positive or negative, are biologically plausible and interesting. Consequently, two-tailed tests are used to obtain the probability of getting an estimate larger in *magnitude* than the one that was obtained had the true value been zero. The two tailed probability is simply the sum of these one-tailed probabilities.

You might not have come across the use of a scaled mean-shifted t-distribution before.

# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

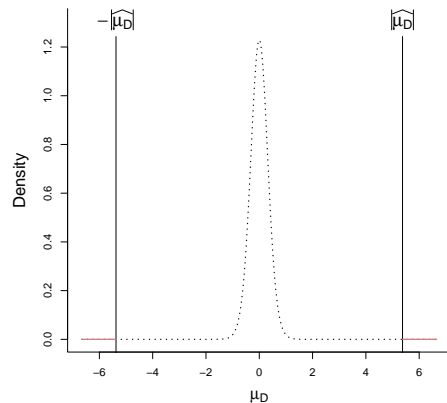
## One-tailed t-test

```
> 1 - pt.scaled(mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```

```
> pt.scaled(-mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.274995e-14
```

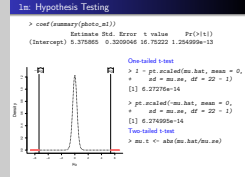
## Two-tailed t-test

```
> mu.t <- abs(mu.hat/mu.se)
```



## Lecture 1

### 1m: Hypothesis Testing



This is because people usually work with a t-value; the absolute value of the estimate divided by the standard error. Under the null hypothesis (the true parameter value is zero) the t-value is also expected to have a mean of zero (it is not mean shifted) and also, dividing by the standard error results in a standard (unscaled) t-distribution.

# 1m: Hypothesis Testing

```
> coef(summary(photo_m1))
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.375865	0.3209046	16.75222	1.254999e-13

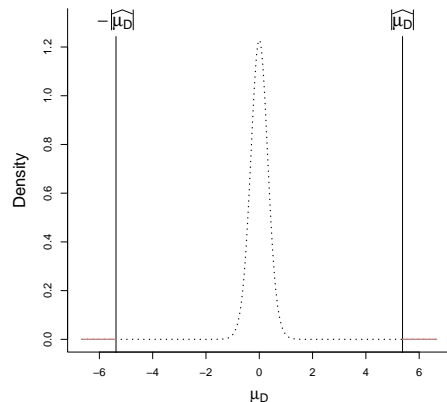
## One-tailed t-test

```
> 1 - pt.scaled(mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.27276e-14
```

```
> pt.scaled(-mu.hat, mean = 0,  
+ sd = mu.se, df = 22 - 1)  
[1] 6.274995e-14
```

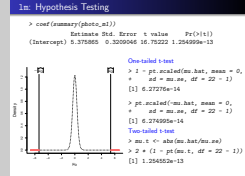
## Two-tailed t-test

```
> mu.t <- abs(mu.hat/mu.se)  
> 2 * (1 - pt(mu.t, df = 22 - 1))  
[1] 1.254552e-13
```



## Lecture 1

### 1m: Hypothesis Testing



We can then just use the cumulative density function of the standard t-distribution.



### └ 1m: Hypothesis Testing

When the data are Gaussian and independent (conditional on the predictors) the assumptions of 1m are met and the t-test is exact. However, in many circumstances we do not know what the sampling distribution is, not even under the null hypothesis. As a consequence we need approximate alternatives that work well.

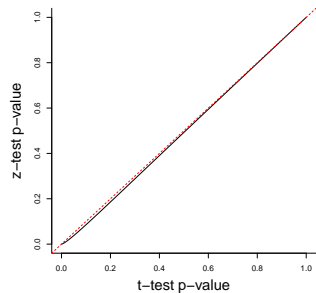
## └ 1m: Hypothesis Testing

### z-test

```
> 2 * (1 - pnorm(mu.t))
```

Replacing the t distribution with a normal distribution, as we did when first calculating the confidence intervals by hand, is one possibility and in many cases it works well. In the case of null-hypothesis testing it is called a z-test. Here we take our t-value (although now we call it a z-value to confuse things) and we ask how likely is it to get a z-value this large or larger in magnitude if the null hypothesis is true. Again, because our null hypothesis is a value of zero the sampling distribution has a mean of zero, and because we have divided through by the standard error (the scale of the sampling distribution) the variance is one. In this case we don't need to specify the mean and variance in the call to `pnorm` because these are the default values. The normal distribution with a mean of zero and variance (and standard deviation) of one is known as the unit normal.

# 1m: Hypothesis Testing



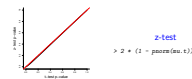
z-test

$$> 2 * (1 - pnorm(mu.t))$$

## Lecture 1

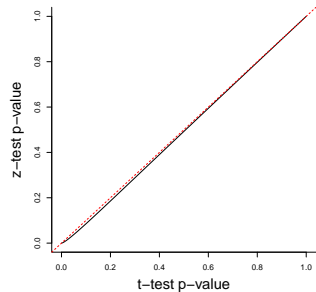
### 1m: Hypothesis Testing

You can see that the normal approximation to the sampling distribution also works well in a hypothesis testing framework. Here I've taken a range of t-values from around 5 down to zero and calculated the p-value using a t-test (x-axis) and a z-test (y-axis) and plotted them in black. The z-test is slightly anti-conservative - the red dashed line is the 1:1 line - and so for a given p-value obtained under the t-test, the p-value under the z-test is always smaller. But the difference is small even given the fact that the size of our data set is modest (21 degrees of freedom).



z-test  
> 2 \* (1 - pnorm(mu.t))

# 1m: Hypothesis Testing



z-test

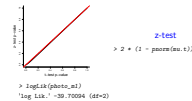
$$> 2 * (1 - pnorm(mu.t))$$

```
> logLik(photo_m1)
'log Lik.' -39.70094 (df=2)
```

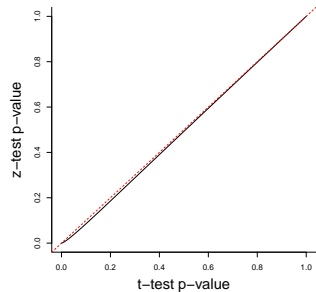
## Lecture 1

### 1m: Hypothesis Testing

1m: Hypothesis Testing



Another possibility is to consider the likelihood of the data. This is the log-likelihood of the model, and you can see it also prints the degrees of freedom. This degrees of freedom is different from the one you saw before. It is simply the number of parameters in the model, both location parameters (the intercept in this model) and also the variance parameters (the residual variance in this model).



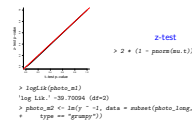
z-test

```
> 2 * (1 - pnorm(mu.t))
```

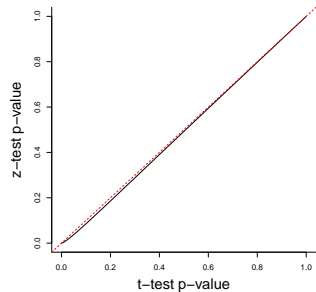
```
> logLik(photo_m1)
'log Lik.' -39.70094 (df=2)
> photo_m2 <- lm(y ~ -1, data = subset(photo_long,
+   type == "grumpy"))
```

### 1m: Hypothesis Testing

What we could do is fit our null model to our data. In this case our null model is that the mean is exactly zero. In R you can specify that you do not want an intercept by 'removing' the intercept using a minus sign.



# 1m: Hypothesis Testing



z-test

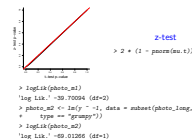
$$> 2 * (1 - pnorm(mu.t))$$

```
> logLik(photo_m1)
'log Lik.' -39.70094 (df=2)
> photo_m2 <- lm(y ~ -1, data = subset(photo_long,
+   type == "grumpy"))
> logLik(photo_m2)
'log Lik.' -69.01266 (df=1)
```

## Lecture 1

### 1m: Hypothesis Testing

1m: Hypothesis Testing



We can then get the log-likelihood of this new model. We can see that the likelihood under our new, null, model is considerably less. The question is, is it significantly less? As we noted earlier, as you increase the number of parameters the likelihood has to go up. In this instance we are comparing a two parameter model with a one parameter model (the variance is still being estimated) and so the question is how much do we expect the likelihood to go up by just by chance?

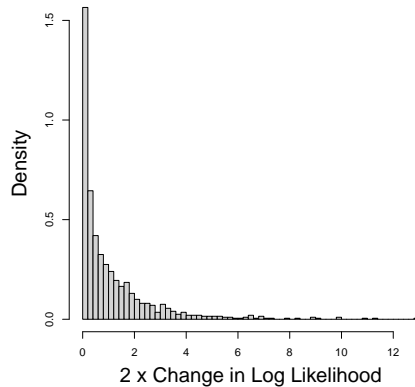
## Likelihood Ratio Test

```
> LR2 <- 2 * (logLik(photo_m1) -  
+           logLik(photo_m2))
```

### └ 1m: Hypothesis Testing

Understanding how much the likelihood should increase just by chance is the basis for something called the likelihood ratio test. The main quantity is the change in the log-likelihood between the two models doubled.

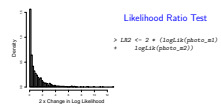
```
> LR2 <- 2 * (logLik(photo_m1) -  
+           logLik(photo_m2))
```



## Likelihood Ratio Test

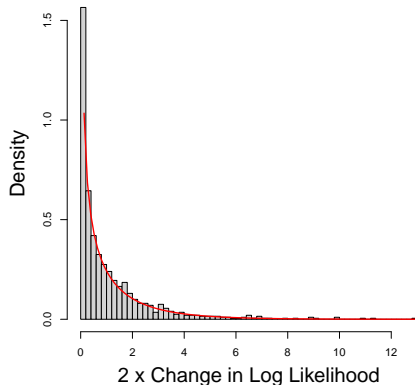
```
> LR2 <- 2 * (logLik(photo_m1) -  
+           logLik(photo_m2))
```

## 1m: Hypothesis Testing



What I have done to generate this figure is simulated data according to our null model (`photo_m2`) where the mean is 0 and the residual variance was estimated to be 31.1 (the variance is so large because the residuals are now the difference between the data and 0 rather than their sample mean). I've then refitted the null model to the simulated data set and extracted the log-likelihood, and then refitted the alternative model (`photo_m1`) to the data set (where the estimate of the mean is allowed to be different from zero) and calculated twice the difference in the log-likelihood. I've done this for many simulated data sets and presented the results as a histogram. In most cases the log-likelihood is changing by less than one unit, but there is a reasonable probability of getting changes in the log-likelihood of 2 or more (i.e.  $LR2 > 4$ ).



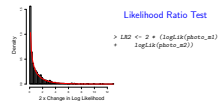


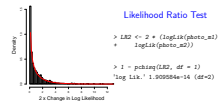
## Likelihood Ratio Test

```
> LR2 <- 2 * (logLik(photo_m1) -
+           logLik(photo_m2))
```

## 1m: Hypothesis Testing

This red line is the density of a chi-squared distribution with one degree of freedom. I've used one degree of freedom because that's the difference in the number of parameters between the models. You can see it seems to fit our empirical distribution rather well, and actually as the information in the data about the parameter goes up (as sample size increases) we know that this distribution will really follow a chi-squared distribution. We can then ask how likely are we to see the increase in likelihood that we actually saw had the underlying mean actually been zero?





## 1m: Hypothesis Testing

Again we can take our observed value ( $LR_2=58.6$ ) and ask what is the probability that the change in likelihood exceeds this? Note that this is a one-tailed test. If we add more parameters to the model the likelihood has to go up, so we are only interested in whether it is possible to get a change in likelihood *greater* than what we observed.

## Likelihood Ratio Test

```

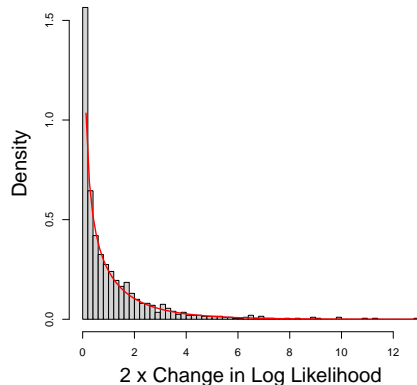
> LR2 <- 2 * (logLik(photo_m1) -
+   logLik(photo_m2))

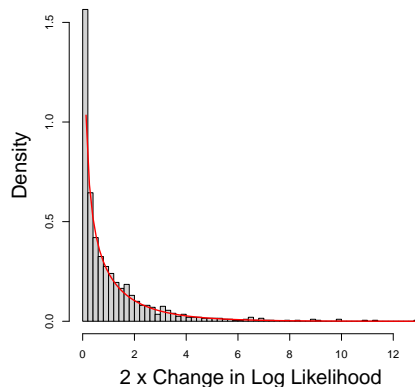
```

```

> 1 - pchisq(LR2, df = 1)
'log Lik.' 1.909584e-14 (df=2)

```





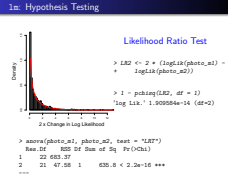
## Likelihood Ratio Test

```
> LR2 <- 2 * (logLik(photo_m1) -
+           logLik(photo_m2))
```

```
> 1 - pchisq(LR2, df = 1)
'log Lik.' 1.909584e-14 (df=2)
```

```
> anova(photo_m1, photo_m2, test = "LRT")
  Res.Df    RSS Df Sum of Sq  Pr(>Chi)
1      22 683.37
2      21  47.58  1    635.8 < 2.2e-16 ***
---
```

## 1m: Hypothesis Testing



We can do this in R by comparing our models using `anova` and specifying `test="LRT"`. There's actually a slight discrepancy here. We haven't really got the time to understand why this discrepancy exists, and for practical purposes I don't think it is important, so I'm going to skip it.

### Summary

- Distributions
  - Data Distribution - probability of data.
  - Sampling Distribution - probability of estimates.
  - Posterior Distribution - probability (epistemic) of parameter values.
- Distribution Functions
  - Mass/Density - (proportional to the) probability that  $X = x$ .
  - Cumulative Density/Mass - probability that  $X \leq x$ .
  - Quantile - Opposite of Cumulative: return probability given  $x$ .
- Inference
  - Maximum Likelihood (ML) - choose parameters that maximise the probability of the data given the model.
  - If the data are Gaussian we have the linear model.
    - Sampling distribution of location parameters are t-distributed (close to Gaussian in many cases).
    - Hypothesis testing with t-test (close to z-test in many cases).
  - Likelihood ratio test for general hypothesis testing under ML.

- Distributions
  - Data Distribution - probability of data.
  - Sampling Distribution - probability of estimates.
  - Posterior Distribution - probability (epistemic) of parameter values.
- Distribution Functions
  - Mass/Density - (proportional to the) probability that  $X = x$ .
  - Cumulative Density/Mass - probability that  $X \leq x$ .
  - Quantile - Opposite of Cumulative: return probability given  $x$ .
- Inference
  - Maximum Likelihood (ML) - choose parameters that maximise the probability of the data given the model.
  - If the data are Gaussian we have the linear model.
    - Sampling distribution of location parameters are t-distributed (close to Gaussian in many cases).
    - Hypothesis testing with t-test (close to z-test in many cases).
  - Likelihood ratio test for general hypothesis testing under ML.