

## Week 4: The Normal Distribution

# Recap

So far we have

- ▶ learned about maximising the likelihood
- ▶ estimated confidence intervals and standard errors

These are the basic tools we will use to fit and understand our models

## More than one datum

So far we have only used one data point. But what if we have more?

If we make one assumption, the maths is easy

## More than one datum: the probability

If data are independent, then

$$Pr(X_1 \& X_2) = Pr(X_1)Pr(X_2)$$

So we can multiply the probabilities

In general, then

$$Pr(X_1, X_2, \dots, X_n) = \prod_{i=1}^n Pr(X_i)$$

## More than one datum: the likelihood

The log-likelihood for the parameters ( $\theta$ ) given the data is

$$l(\theta|X_1, X_2, \dots, X_n) = \sum_{i=1}^n \log(\text{Pr}(X_i|\theta))$$

So we just add the log-likelihoods together

- ▶ easier than multiplying!

## Punxsutawney Phil & Groundhog Day

Here is some data on whether Punxsutawney Phil predicts another 6 weeks of winter (which he tries to do every Feb 2)

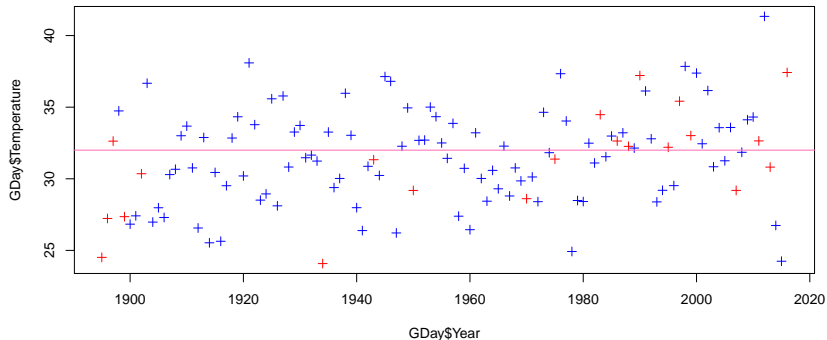
We will look at the average temperature for Feb/March in Pennsylvania in each year

```
GDay <- read.csv(file="https://www.math.ntnu.no/emner/ST230")
```

# Winter/Spring Temperatures

These are the mean February/March temperatures

```
plot(GDay$Year, GDay$Temperature, cex=1, pch=3,  
     col=c("red", "blue")[1+GDay$Shadow])  
abline(h=32, col="hotpink")
```



Initial question: is the average temperature  $> 32$ F (i.e. freezing)?

# The Normal Distribution

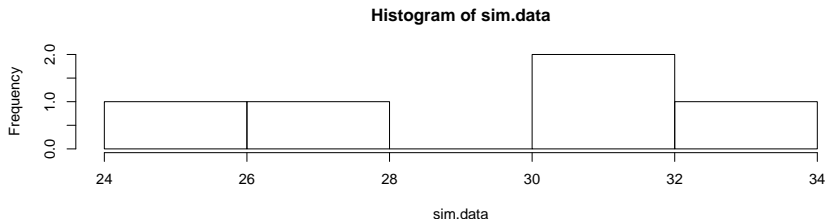
$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



# The Normal Distribution: Exercises I

Simulate 100 data points from a normal distribution with the same mean and standard deviation as the temperature data, using `rnorm()`

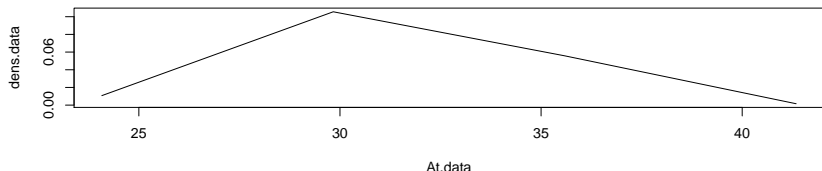
```
MeanTemp <- mean(GDay$Temperature)
sdTemp <- sd(GDay$Temperature)
sim.data <- rnorm(5, mean = MeanTemp, sd = sdTemp)
hist(sim.data)
```



## The Normal Distribution: Exercises II

- ▶ Calculate the density data from the distribution for values using `dnorm()`
  - ▶ why is the density different from the probability (hint: the normal is continuous)

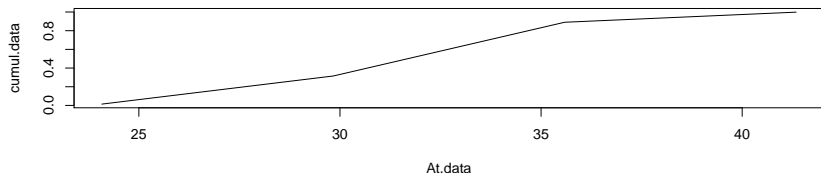
```
At.data <- seq(min(GDay$Temperature),  
              max(GDay$Temperature), length=4)  
dens.data <- dnorm(At.data, mean = MeanTemp, sd = sdTemp)  
plot(At.data, dens.data, type="l")
```



## The Normal Distribution: Exercises III

- ▶ Calculate the cumulative density data from the distribution for values using `pnorm()`

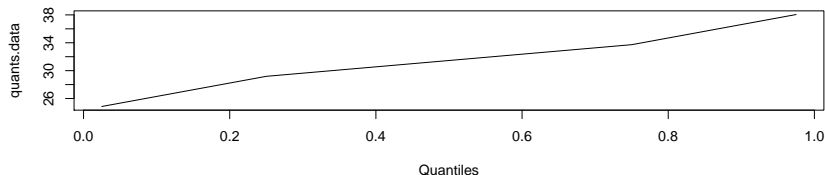
```
cumul.data <- pnorm(At.data, mean = MeanTemp, sd = sdTemp)  
plot(At.data, cumul.data, type="l")
```



## The Normal Distribution: Exercises IV

- ▶ Calculate the 2.5th, 25, 50th, 75th, and 97.5th quantiles from the distribution using `qnorm()`

```
Quantiles <- c(0.025, 0.25, 0.5, 0.75, 0.975)
quants.data <- qnorm(Quantiles,
                    mean = MeanTemp, sd = sdTemp)
plot(Quantiles, quants.data, type="l")
```



What is the difference between `pnorm()` and `qnorm()`?

## The Normal Distribution: the log likelihood

If we take logs of the density,

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

we get this:

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

The likelihood for a single data point. For  $\mu$  this is just a quadratic

## The Likelihood for $\mu$

The likelihood for  $n$  independent samples is the product of each likelihood:

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

This means that the log-likelihood is the sum of the likelihoods, the sum of quadratic terms:

$$\log L(\mu, \sigma | x_1, \dots, x_n) = \sum_{i=1}^n l(x_i | \mu, \sigma) = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

In practice, we can calculate this with `dnorm(..., log=TRUE)`

# Our Task

Estimate the parameters of this distribution

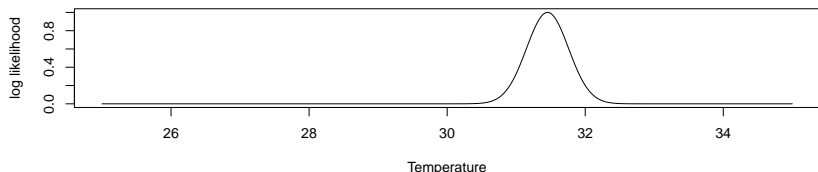
- ▶ estimate  $\hat{\mu}$  and  $\hat{\sigma}^2$

In practice,  $\hat{\mu}$  is more important, because we will be modelling  $\mu$  as a function of different effects

## Finding the Estimate

We can simulate data and calculate the likelihood for different values of the mean (we will fix the standard deviation for now)

```
CalcNormLh <- function(mu, sigma, data) {  
  lhood <- sum(dnorm(data, mean=mu, sd=sigma, log=TRUE))  
  lhood  
}  
Means <- seq(25, 35, length=500)  
lhoods <- sapply(Means, CalcNormLh, sigma=sdTemp,  
                 data=GDay$Temperature)  
plot(Means, exp(lhoods-max(lhoods)), type="l",  
      xlab="Temperature", ylab = "log likelihood")
```



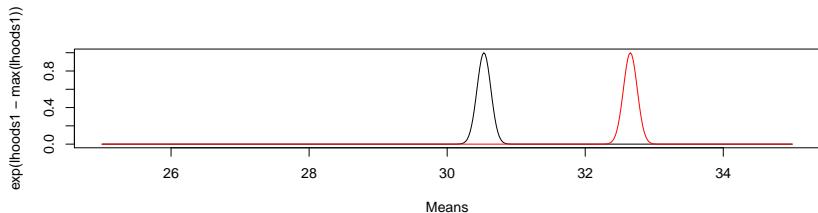


## Plotting Several Curves

You can plot several curves on the same plot by plotting one, and then using `lines()` to add another line to the first plot, e.g.

```
D1 <- rgamma(1e2,60,2); D2 <- rgamma(1e2,65,2)

lhoods1 <- sapply(Means, CalcNormLh, sigma=1, data=D1)
lhoods2 <- sapply(Means, CalcNormLh, sigma=1, data=D2)
plot(Means, exp(lhoods1-max(lhoods1)), type="l")
lines(Means, exp(lhoods2-max(lhoods2)), col=2)
```



## Different Amounts of Data

Simulate data with the same mean & standard deviation as the Groundhog Day data and plot the likelihood for

- ▶ 10 data points
- ▶ 100 data points
- ▶ 1000 data points

Fine (approximately!) the maximum likelihood estimate

What happens to the likelihood when you have more data?

## The MLE for $\mu$

We could try simulating & finding the best value, or we could try numerically maximising this. But we can get an analytic solution (this is one reason why the normal distribution is so nice - the maths is relatively easy)

## The MLE for $\mu$

We can differentiate the log-likelihood w.r.t  $\mu$ , and set this to zero

$$0 = \frac{1}{2\sigma^2} \left( 2 \sum_{i=1}^n x_i - 2n\mu \right)$$

Then re-arrange, and the MLE is

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

The sample mean!

## The MLE for $\sigma^2$

This is usually less important. We are generally not interested in the standard deviation, but it is a parameter of the distribution, so it has to be estimated. What we do is differentiate w.r.t  $\sigma^2$ , set to zero, re-arrange, and get

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

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

## Comments

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

- ▶ the whole distribution can be summarised by these two statistics

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

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

## The distribution of $\hat{\mu}$

We can look at the distribution of  $\hat{\mu}$ , and (for example) estimate confidence intervals:

## The distribution of $\hat{\mu}$

If the data are normally distributed, the distribution of  $\hat{\mu}$  is a t-distribution.

- ▶ we can use `dt`, `pt` etc.

The parameters:

- ▶ mean
- ▶ standard error (= standard deviation/ $\sqrt{n}$ )
- ▶ degrees of freedom (=  $n - 1$ )

```
dt(32, mean(GDay$Temperature), sd(GDay$Temperature)/sqrt(10),  
   df=length(GDay$Temperature)-1)
```

```
## [1] 0.1690379
```

If we have enough data, this distribution look like a normal distribution



## Your work

Use `pt()` to calculate the probability that the mean temperature is less than 32F

Use `qt()` to calculate the 95% confidence interval for the mean, i.e. calculate the 2.5% and 97.5% quantiles

- ▶ you can adapt the code for `pnorm()` and `qnorm()`

Next: something useful...

So far we have been learning about statistical inference and statistical programming. Now we can start to use this in modelling.

Next, we will start with a simple model, but put it in the context of maximum likelihood

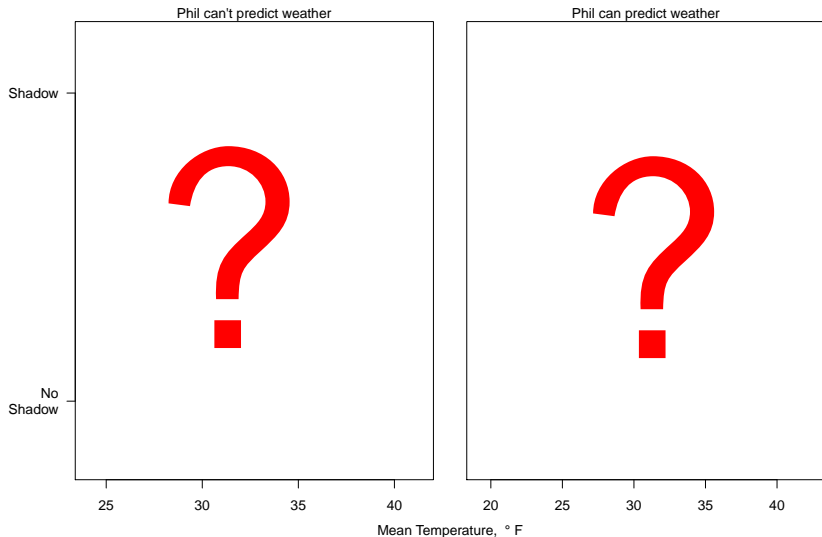
## Predicting the End of Winter

If Punxsutawney Phil sees his shadow, there will be 6 more weeks of winter

If he is good at predicting winter, we should see lower average temperatures in the 2 months after the prediction

# The Modelling

First, what would we expect if Punxsutawney Phil can predict winter, and if he cannot



# The Model

The question is about the mean temperature: is there a clear difference when Phil sees his shadow or not?

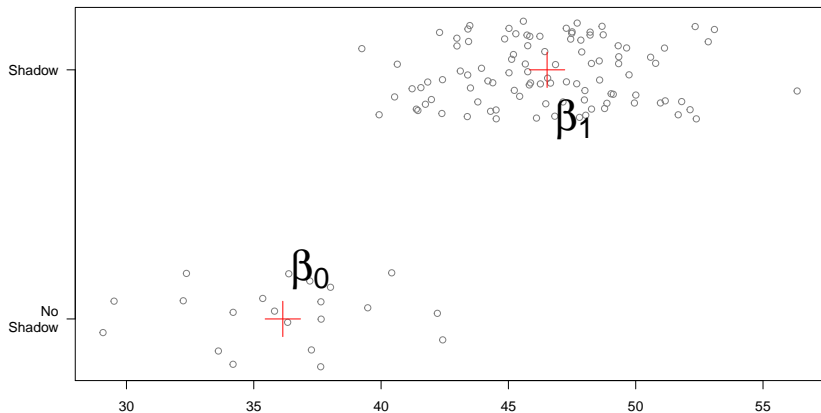
So, we have to build a model where there are different mean temperatures when he does see his shadow, and when he does not.

We can do this in a few ways. For all we assume  $y_i \sim N(\mu_i, \sigma^2)$  (i.e. the data follow a normal distribution with the same variance)

We also will define a variable  $X_i$ :  $X_i = 1$  if Phil saw his shadow (and hence predicted winter),  $X_i = 0$  if he did not

# The Model - First way

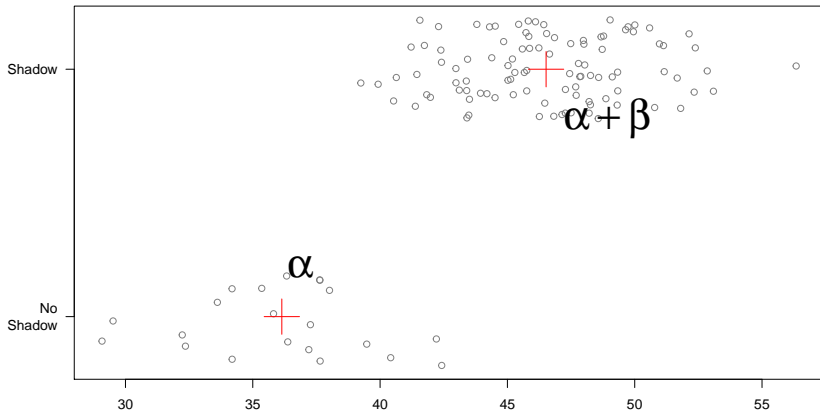
$$\mu_i = \begin{cases} \beta_0 & \text{if } X_i = 0 \\ \beta_1 & \text{if } X_i = 1 \end{cases}$$



## The Model - Second way

$$\mu_i = \begin{cases} \alpha & \text{if } X_i = 0 \\ \alpha + \beta & \text{if } X_i = 1 \end{cases}$$

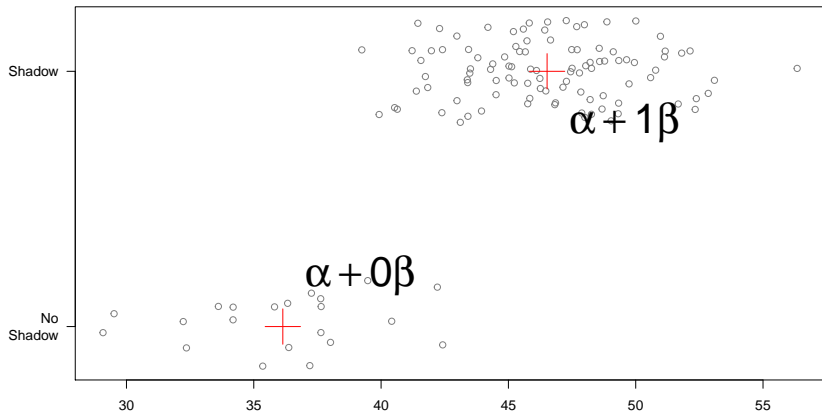
So the difference is  $\beta$ , and this is what we are interested in



## The Model - Third way

$$\mu_i = \alpha + \beta X_i$$

The difference is  $\beta$  (because  $X_i$  can only be 0 or 1). This approach is the easiest to extend to more complex models





## Calculating the Likelihood

The log-likelihood is not too difficult to write down

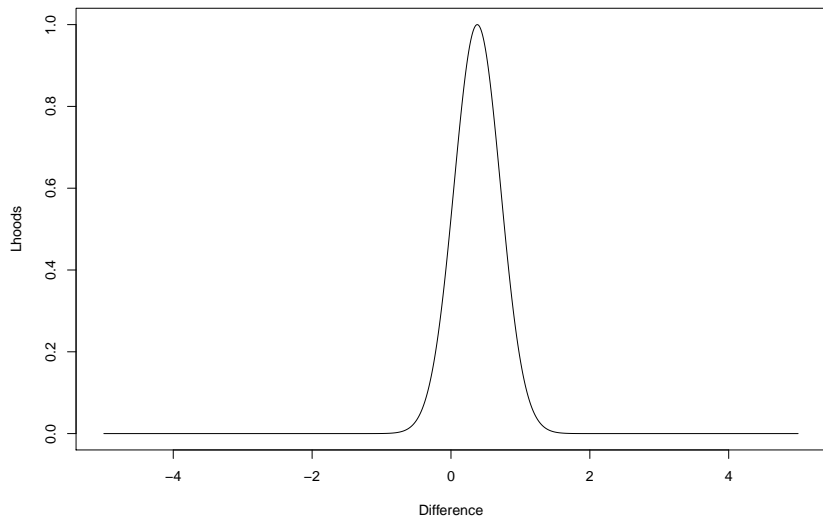
$$\log L(\mu_i, \sigma | x_1, \dots, x_n) = C - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_i)^2$$

where  $\mu_i$  is described in the previous slide.

We now have 3 parameters (two for the means, and the standard deviation). I will not ask you to calculate the likelihood in all 3 dimensions.

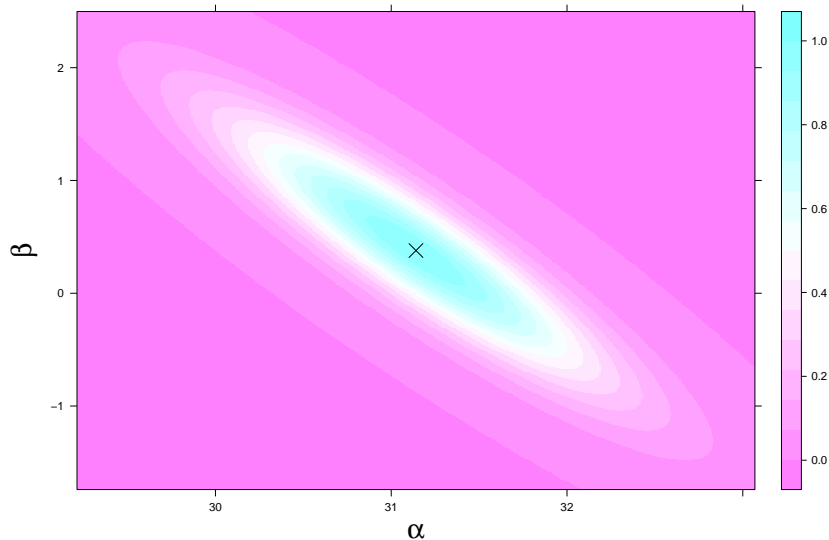
## Calculate the likelihood for the difference

We can look at the likelihood for the difference, using the MLEs for the other parameters



## Estimating the Likelihood for the difference

Even if we ignore  $\sigma$ , the likelihood is more complicated. This is a plot of the likelihood for  $\mu_1$  and  $\beta$



## In Practice

There is much more about this in Chapter 3 of *A New Stats with R*

R has functions to do the hard work that we have been doing

For the t-test we can use `t.test()`, but we're going to do something different.

We can use Model 3 ( $\mu_i = \alpha + \beta X_i$ ) directly

## In R

R has a function, `lm()`. We tell R what to use like this:

```
Winter <- lm(GDay$Temperature ~ GDay$Shadow)
# Neater...
Winter <- lm(Temperature ~ Shadow, data=GDay)
```

The model is described as  $Y \sim X$

- ▶ Y is the *response*
- ▶ X is the *explanatory variable* (here it is whether Phil sees his shadow)
- ▶ this is over-kill now, but later in the course you will see that it helps

## Estimates and Confidence intervals

We can extract the estimates with `coef()`

```
coef(Winter)
```

```
## (Intercept)      Shadow  
##  31.1406780    0.3786877
```

This gives two estimates: (Intercept) is for  $\alpha$ , Shadow is for  $\beta$

We get confidence intervals with `confint()`

```
confint(Winter)
```

## Your Tasks

- ▶ what is the estimate of the difference in temperature between when Phil sees his shadow and when he does not?
- ▶ what is the confidence interval?
- ▶ how sure can you be about the direction of any difference? And how large could a difference be?
- ▶ do you think Punxsutawney Phil can predict whether there will be another 6 weeks of winter?