Institut de recherche
pour le développement

Estimation Methods

With an Emphasis on Maximum Likelihood

Emmanuel Paradis

Institut Pertanian Bogor

October 10, 2011

Biological phenomena are characterized by variability at many levels. For instance:

- body growth,
- survial and fecundity (fitness),
- mutation rates, . . .

vary

- through space,
- with time,
- within species (or population),
- among species, . . .

Variability:

- ***deterministic*** (systematic, “explained”)
- ***stochastic*** (probabilistic, random, “error”, “unexplained”)

Variability:

- ***deterministic*** (systematic, “explained”)
- ***stochastic*** (probabilistic, random, “error”, “unexplained”)

Statistical data analysis is building models to quantify (and ultimately predict) the variability we observe in natural phenomena.

Probability has a long history in gambling (early 17th century).

Scientific applications in physics (e.g., calculation of planet trajectories) in the 18th century.

Appeared in biology in the early 20th century.

The ***probability density function***, $f(x)$, often called “density” or “pdf”.

The ***probability density function***, $f(x)$, often called “density” or “pdf”.

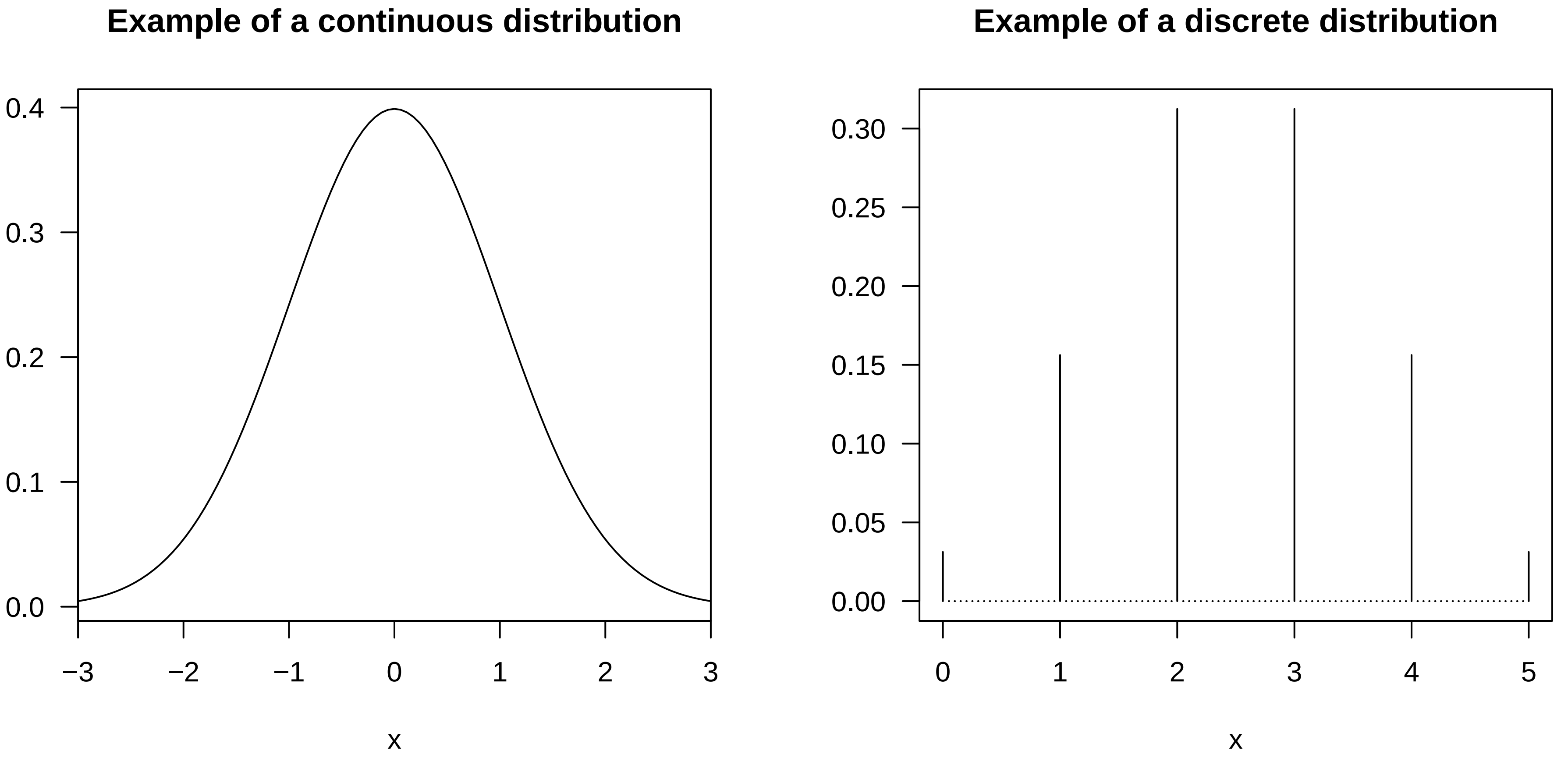
The ***cumulative probability density function*** (often called “CDF”) is the probability that X is smaller or equal to x, i.e., $\mathsf{Pr}(X \leq x)$.

The ***probability density function***, $f(x)$, often called “density” or “pdf”.

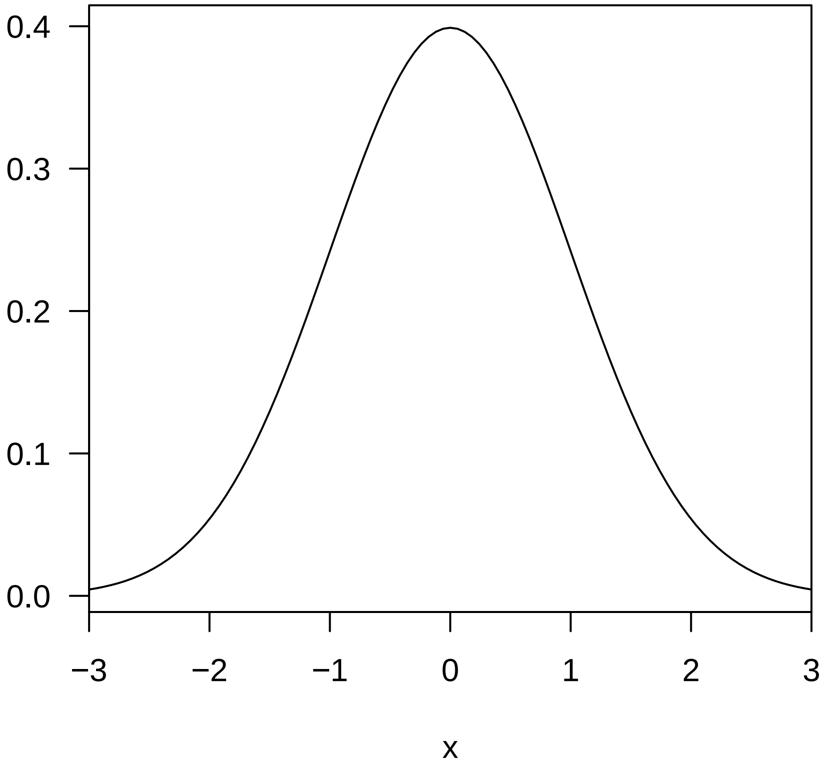
The ***cumulative probability density function*** (often called “CDF”) is the probability that X is smaller or equal to x, i.e., $\mathsf{Pr}(X \leq x)$.

$F(x) = \int_{-\infty}^{x} f(u)\mathsf{d}u$ if x is continuous
$F(x) = \sum_{-\infty}^{x} f(u)$ if x is discrete

X: the variable (abstract)
x: the observation (concrete)

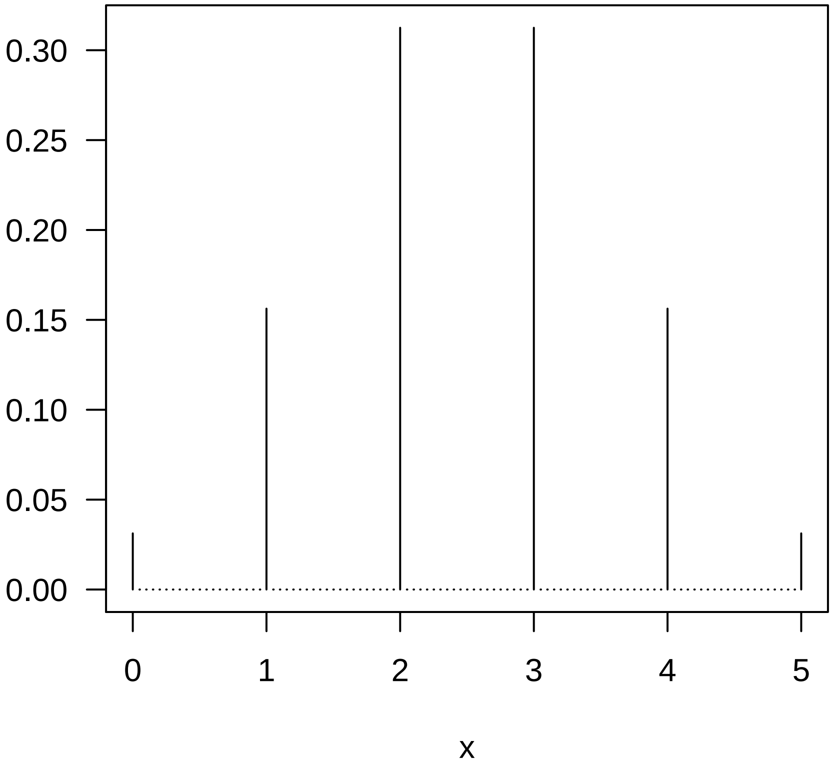

Example of a continuous distribution
0.4
0.3
0.2
0.1
0.0
−3
−2
−1
0
1
2
3
x
Example of a discrete distribution
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
4
5
x

Example of a continuous distribution



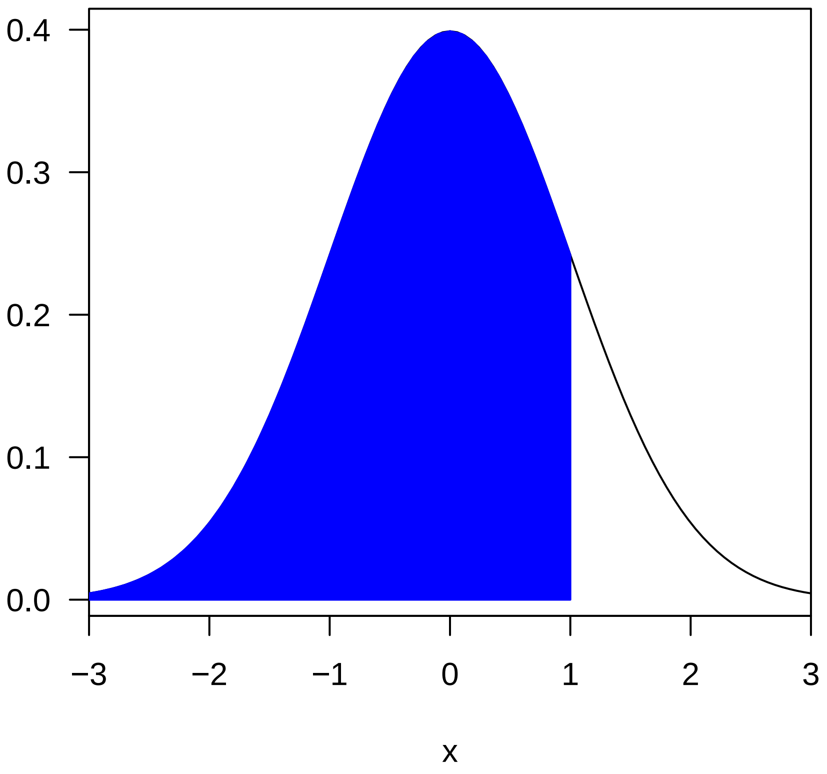
$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$$

Example of a discrete distribution

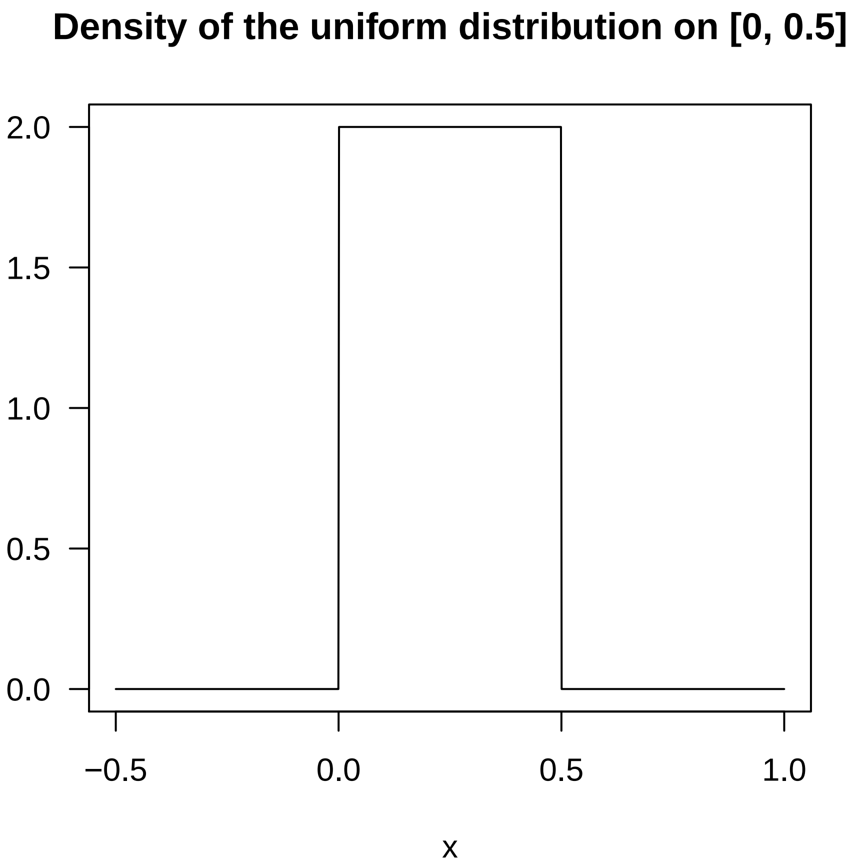


$$\sum_{-\infty}^{\infty} f(x) = 1$$

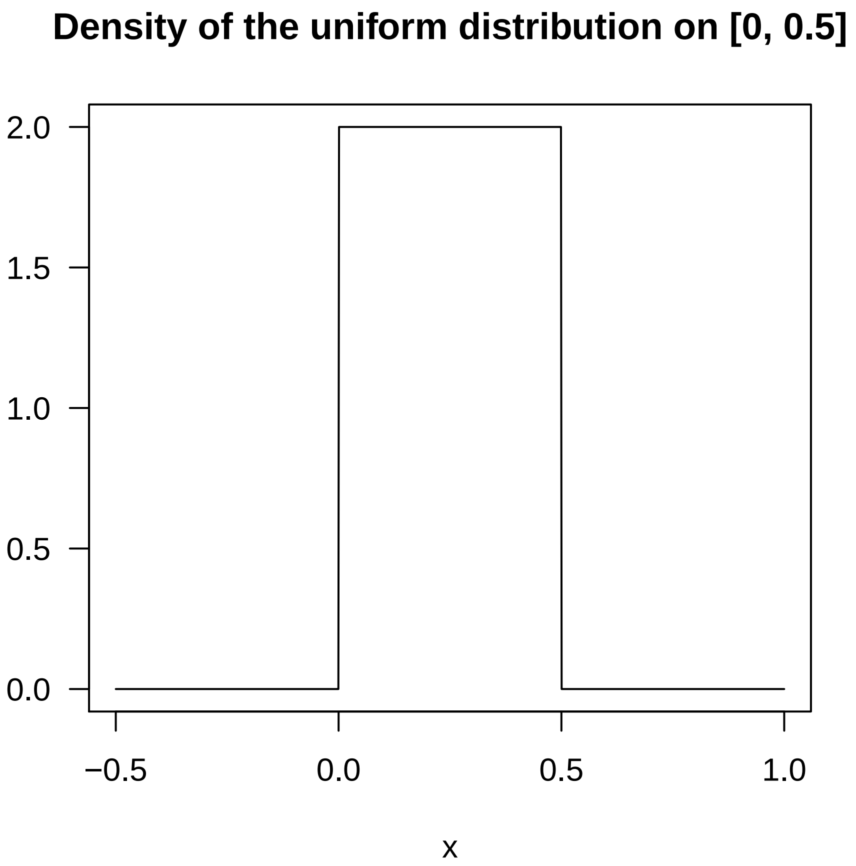
The CDF of the normal distribution with $\mu = 0$ and $\sigma = 1$:



For a continuous variable, the density $f(x)$ is ***not*** the probability $\mathtt{Pr}(X = x)$.



For a continuous variable, the density $f(x)$ is ***not*** the probability $\Pr(X = x)$.



But still $\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1$.

The probability is given by the CDF (or integrating the pdf over a small interval):

$$\text{Pr}(x < X < x + \delta)$$

The probability is given by the CDF (or integrating the pdf over a small interval):

$$\mathsf{Pr}(x < X < x + \delta) = \int_x^{x+\delta} f(u)\mathsf{u}$$

The probability is given by the CDF (or integrating the pdf over a small interval):

$$\mathsf{Pr}(x < X < x + \delta) = \int_{x}^{x+\delta} f(u)\mathsf{u} = F(x + \delta) - F(x)$$

The probability is given by the CDF (or integrating the pdf over a small interval):

$$\mathsf{Pr}(x < X < x + \delta) = \int_x^{x+\delta} f(u)\mathsf{u} = F(x + \delta) - F(x)$$

In our example with the uniform distribution $\mathsf{Pr}(x < X < x + \delta) = 2 \times \delta$.

The probability is given by the CDF (or integrating the pdf over a small interval):

$$\mathsf{Pr}(x < X < x + \delta) = \int_{x}^{x+\delta} f(u)\mathsf{u} = F(x + \delta) - F(x)$$

In our example with the uniform distribution $\mathsf{Pr}(x < X < x + \delta) = 2 \times \delta$. Since $\delta \leq 0.5$, this quantity cannot exceed 1, so it is effectively a probability.

The probability is given by the CDF (or integrating the pdf over a small interval):

$$\mathsf{Pr}(x < X < x + \delta) = \int_{x}^{x+\delta} f(u)\mathsf{u} = F(x + \delta) - F(x)$$

In our example with the uniform distribution $\mathsf{Pr}(x < X < x + \delta) = 2 \times \delta$. Since $\delta \leq 0.5$, this quantity cannot exceed 1, so it is effectively a probability.

However, we note the proportionality $\mathsf{Pr}(X = x) \propto f(x)$.

The probability is given by the CDF (or integrating the pdf over a small interval):

$$\text{Pr}(x < X < x + \delta) = \int_{x}^{x+\delta} f(u)\text{u} = F(x + \delta) - F(x)$$

In our example with the uniform distribution $\text{Pr}(x < X < x + \delta) = 2 \times \delta$.
Since $\delta \leq 0.5$, this quantity cannot exceed 1, so it is effectively a probability.

However, we note the proportionality $\text{Pr}(X = x) \propto f(x)$.

Thousands of distributions have been developed with one or more parameters:

$f_\theta(x)$: density $\qquad$ θ: vector of parameters

If θ is known, f is fully determined.

How to know θ from observations?

An ***estimator*** is a method to estimate parameters.

An ***estimate*** is a value of an estimator computed for a given data set, usually denoted as $\widehat{\theta}$.

How to know θ from observations?

An ***estimator*** is a method to estimate parameters.

An ***estimate*** is a value of an estimator computed for a given data set, usually denoted as $\widehat{\theta}$.

Methods of moments is suitable for simple distributions (e.g., estimating the mean of a normal distribution).

Methods of least squares is for models with a normally distributed stochastic component (e.g., linear regression).

Methods of maximum likelihood is the most general method of statistical estimation.

The idea of *likelihood* is intuitive: we calcule the probability to observe the data.

The idea of *likelihood* is intuitive: we calcule the probability to observe the data. But this quantity cannot be interpreted as a probability because we cannot define a set of events whose sum of probabilities is 1. Therefore, likelihood has no predictive value.

The idea of *likelihood* is intuitive: we calcule the probability to observe the data. But this quantity cannot be interpreted as a probability because we cannot define a set of events whose sum of probabilities is 1. Therefore, likelihood has no predictive value. If the observations are independent:

$$L = \prod_{i=1}^{n} f_\theta(x_i) \qquad n : \text{sample size}$$

The idea of *likelihood* is intuitive: we calcule the probability to observe the data. But this quantity cannot be interpreted as a probability because we cannot define a set of events whose sum of probabilities is 1. Therefore, likelihood has no predictive value. If the observations are independent:

$$L = \prod_{i=1}^{n} f_\theta(x_i) \qquad n : \text{sample size}$$

```
> x <- rnorm(10, mean = 0)
> prod(dnorm(x, mean = 0))
[1] 4.370395e-06
> prod(dnorm(x, mean = -1))
[1] 4.491077e-09
> prod(dnorm(x, mean = 1))
[1] 1.930839e-07
```

```
> x <- rnorm(600)
> prod(dnorm(x))
[1] 0
```

```
> x <- rnorm(600)
> prod(dnorm(x))
[1] 0
```

The log-likelihood: $$\ln L = \sum_{i=1}^{n} \ln f_\theta(x_i)$$

```
> sum(dnorm(x, log = TRUE))
[1] -12.34066
> sum(dnorm(x, -1, log = TRUE))
[1] -19.22117
> sum(dnorm(x, 1, log = TRUE))
[1] -15.46014
```

The principle of ***maximum likelihood estimation*** is to find the value of θ that maximizes the (log-)likelihood function. This value is called the maximum likelihood estimator (MLE) of θ.

If the log-likelihood function is simple, this can be solved analytically with:

$$\frac{\partial \ln L}{\partial \theta} = 0$$

If the log-likelihood function is simple, this can be solved analytically with:

$$\frac{\partial \ln L}{\partial \theta} = 0$$

Example with the exponential distribution: $f(x) = \lambda e^{-\lambda x}$. The likelihood for a sample $x_1, \ldots, x_n$ is:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

If the log-likelihood function is simple, this can be solved analytically with:

$$\frac{\partial \ln L}{\partial \theta} = 0$$

Example with the exponential distribution: $f(x) = \lambda e^{-\lambda x}$. The likelihood for a sample $x_1, \ldots, x_n$ is:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

So the log-likelihood is:

$$\ln L = \sum \ln(\lambda e^{-\lambda x_i})$$

If the log-likelihood function is simple, this can be solved analytically with:

$$\frac{\partial \ln L}{\partial \theta} = 0$$

Example with the exponential distribution: $f(x) = \lambda e^{-\lambda x}$. The likelihood for a sample $x_1, \ldots, x_n$ is:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

So the log-likelihood is:

$$\ln L = \sum \ln(\lambda e^{-\lambda x_i}) = \sum \ln \lambda - \lambda x_i$$

If the log-likelihood function is simple, this can be solved analytically with:

$$\frac{\partial \ln L}{\partial \theta} = 0$$

Example with the exponential distribution: $f(x) = \lambda e^{-\lambda x}$. The likelihood for a sample $x_1, \ldots, x_n$ is:

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

So the log-likelihood is:

$$\ln L = \sum \ln(\lambda e^{-\lambda x_i}) = \sum \ln \lambda - \lambda x_i = n \ln \lambda - \lambda \sum x_i$$

With a partial derivative with respect to λ:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum x_i$$

whose maximum is easily found by solving:

$$\frac{n}{\lambda} - \sum x_i = 0$$

With a partial derivative with respect to λ:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - \sum x_i$$

whose maximum is easily found by solving:

$$\frac{n}{\lambda} - \sum x_i = 0 \qquad \Rightarrow \qquad \hat{\lambda} = \frac{n}{\sum x_i}$$

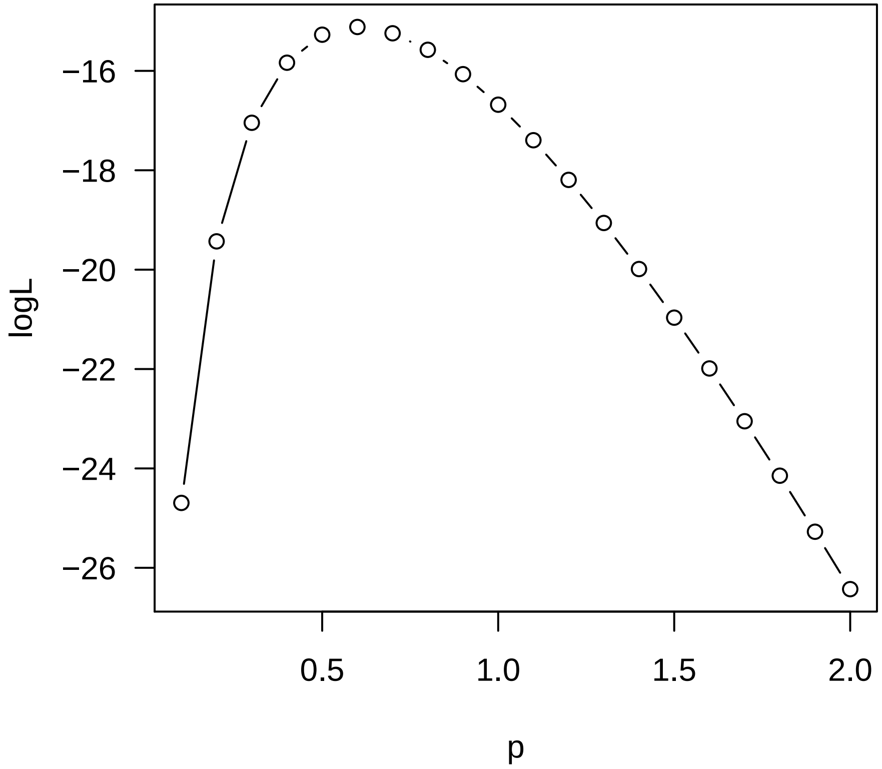
The second derivatives give a confidence interval for the maximum likelihood estimators:

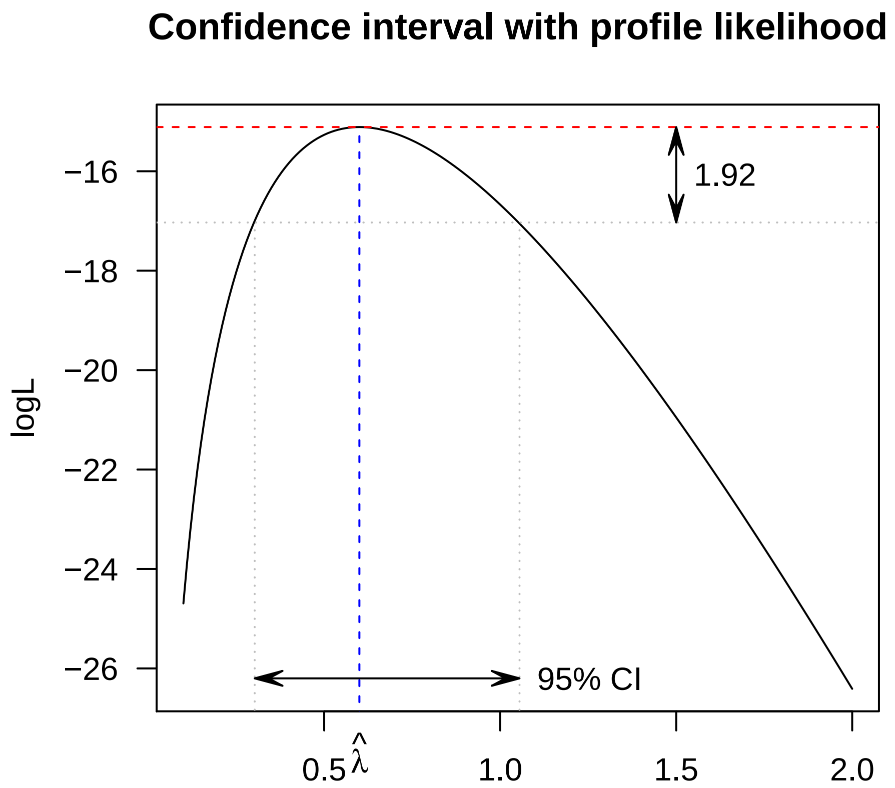
$$\mathsf{SE}(\widehat{\theta}) = \left(\left[-\frac{\partial^2 \ln L}{\partial \theta^2}\right]_{\widehat{\theta}}\right)^{-\frac{1}{2}}$$

In most cases such analytical solutions cannot be found and numerical methods must be used which are very general.

In most cases such analytical solutions cannot be found and numerical methods must be used which are very general. Example with the exponential distribution:

```
x <- rexp(10, 1)
logL <- sum(dexp(x, p, log = TRUE))
```


−16
−18
−20
−22
−24
−26
logL
0.5
1.0
1.5
2.0
p



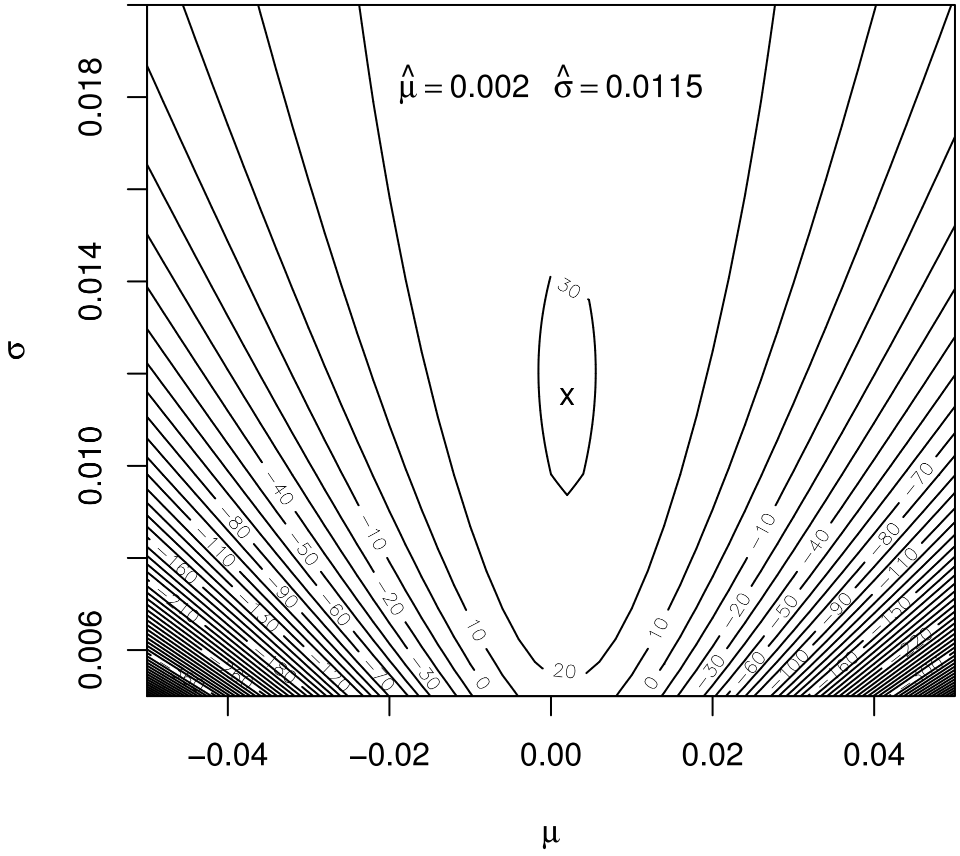
Confidence interval with profile likelihood

With more than one parameter (e.g., the normal distribution):

$$\ln L = \sum_i \ln f_\theta(x_i) \qquad \theta = \{\mu, \sigma^2\}$$

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

$$\frac{\partial \ln L}{\partial \mu} = 0 \qquad \frac{\partial \ln L}{\partial \sigma} = 0$$


$\hat{\mu} = 0.002$ $\hat{\sigma} = 0.0115$
x
30
20
10
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
−100
−110
−120
−130
−150
−160
−210
−220
σ
0.006
0.010
0.014
0.018
μ
−0.04
−0.02
0.00
0.02
0.04

So far, we have considered parameter estimation for a ***given*** model, but how to choose a model among several alternatives?

So far, we have considered parameter estimation for a ***given*** model, but how to choose a model among several alternatives?

If two models are nested (one is a special case of another), the ratio of their likelihood multiplied by 2 follows a χ^2 distribution with a number of degrees of freedom given by the difference in their number of parameters (*likelihood ratio test*, LRT):

$$Dev_1 - Dev_2 \sim \chi^2 \quad df = k_2 - k_1$$

$Dev = -2 \ln L$, k: number of (free) parameters

So far, we have considered parameter estimation for a ***given*** model, but how to choose a model among several alternatives?

If two models are nested (one is a special case of another), the ratio of their likelihood multiplied by 2 follows a χ^2 distribution with a number of degrees of freedom given by the difference in their number of parameters (*likelihood ratio test*, LRT):

$$Dev_1 - Dev_2 \sim \chi^2 \quad df = k_2 - k_1$$

$Dev = -2 \ln L$, k: number of (free) parameters

Ex: $Y = \beta X + \alpha$ and $Y = \alpha$ (see below)

If two models are not nested, they can be compared with the Akaike information criterion $AIC = Dev + 2 \times k$. The model wit the smallest AIC value must be selected.

AIC has no absolute meaning and must used to select models based on the ***same*** data.

AIC has many variants but they often give the same results.

The linear model: $Y = \beta X + \alpha$

The observations: $y_i = \beta x_i + \alpha + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

The linear model: $Y = \beta X + \alpha$

The observations: $y_i = \beta x_i + \alpha + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$\beta x_i + \alpha$: deterministic compoment

ϵ_i: stochastic compoment

The linear model: $Y = \beta X + \alpha$

The observations: $y_i = \beta x_i + \alpha + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$\beta x_i + \alpha$: deterministic compoment
ϵ_i: stochastic compoment

$y_i \sim \mathcal{N}(\beta x_i + \alpha, \sigma^2)$

y_i is normally distributed conditionally on x_i.

The linear model: $Y = \beta X + \alpha$

The observations: $y_i = \beta x_i + \alpha + \epsilon_i \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

$\beta x_i + \alpha$: deterministic compoment

ϵ_i: stochastic compoment

$y_i \sim \mathcal{N}(\beta x_i + \alpha, \sigma^2)$

y_i is normally distributed conditionally on x_i.

$$L = \prod_{i=1}^{n} f_\theta(y_i) \qquad \theta = \{\beta, \alpha, \sigma^2\}$$

```
fd  <- function(p) {
    m <- p[1] * x + p[2]
    -2*sum(dnorm(y, m, p[3], log = TRUE))
}
```

with `p[1]`: β `p[2]`: α `p[3]`: σ^2

```
> x <- 1:50
> y <- 1.5 * x + 8 + rnorm(50, 0, 10)
> nlm(fd, c(1, 1, 1))
$minimum
[1] 379.3798

$estimate
[1]  1.301263 14.872885 10.749500
....
> mod <- lm(y ~ x)
> mod$coeff
(Intercept)           x
  14.872842    1.301265
> summary(mod)$sigma
[1] 10.97117
> AIC(mod)
[1] 385.3798
```

The application of maximum likelihood estimation are extremely vast.

Generalized linear models (GLM): linear models with non-normal “errors”

Mixed effects models: liner models with several random components

Survival models (exponential distribution)

Phylogenetics and molecular evolution

. . .

The relationship between least squares and maximum likelihood

If we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then fitting a model by maximum likelihood (ML) or least squares (LS) is equivalent.

ML: maximize the (log)-likelihood function (see above)
LS: minimize the sums of residuals squares $\sum_i (y_i - \widehat{y}_i)^2$

The relationship between least squares and maximum likelihood

If we assume $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then fitting a model by maximum likelihood (ML) or least squares (LS) is equivalent.

ML: maximize the (log)-likelihood function (see above)
LS: minimize the sums of residuals squares $\sum_i (y_i - \widehat{y}_i)^2$

```
> x <- 1:5
> y <- c(7.7, 8.1, 24.9, 33.2, 13.3)
> summary(glm(y ~ x))
....
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.550     11.671   0.561    0.614
```

```
x               3.630      3.519   1.032    0.378

(Dispersion parameter for gaussian family taken to be 123.8343)
....
> summary(lm(y ˜ x))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.550     11.671   0.561    0.614
x              3.630      3.519   1.032    0.378

Residual standard error: 11.13 on 3 degrees of freedom
....
> 11.13ˆ2
[1] 123.8769
> AIC(lm(y ˜ x))
```

```
[1] 41.72998
> AIC(glm(y ~ x))
[1] 41.72998
```

This applies also to analysis of variance (ANOVA) and analysis of covariance (ANCOVA).

A case where least squares cannot be reasonably used

```
> x <- 1:5
> y <- c(0, 1, 0, 1, 1)
```

A case where least squares cannot be reasonably used

```
> x <- 1:5
> y <- c(0, 1, 0, 1, 1)
```

We assume that y follows a binomial distribution: the probability p will be transformed to use a linear model:

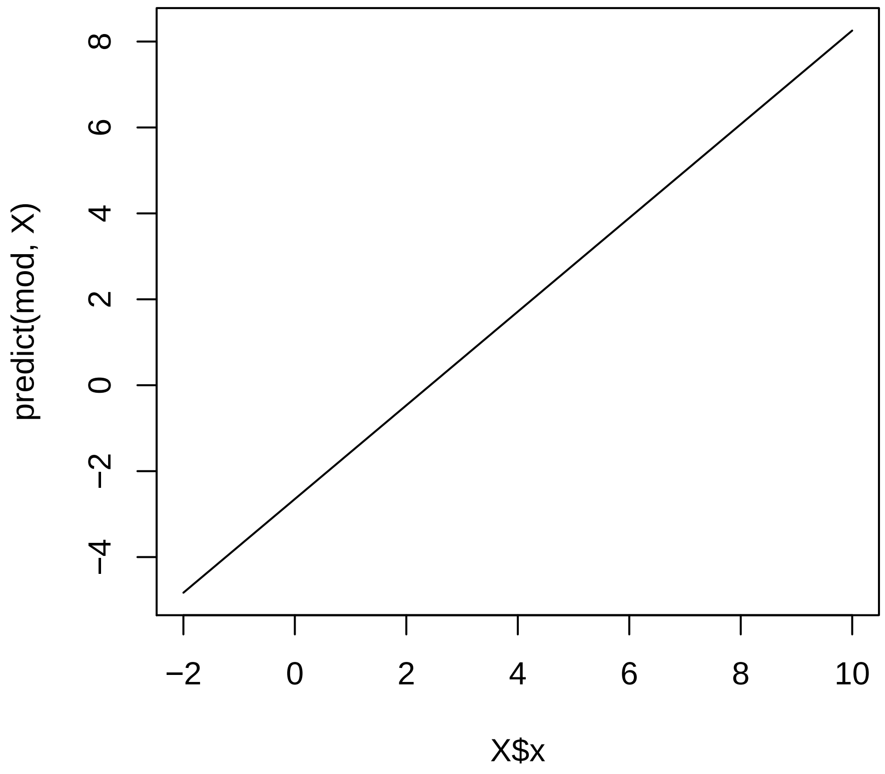
$$\ln\left(\frac{p}{1-p}\right) = \beta X + \alpha$$

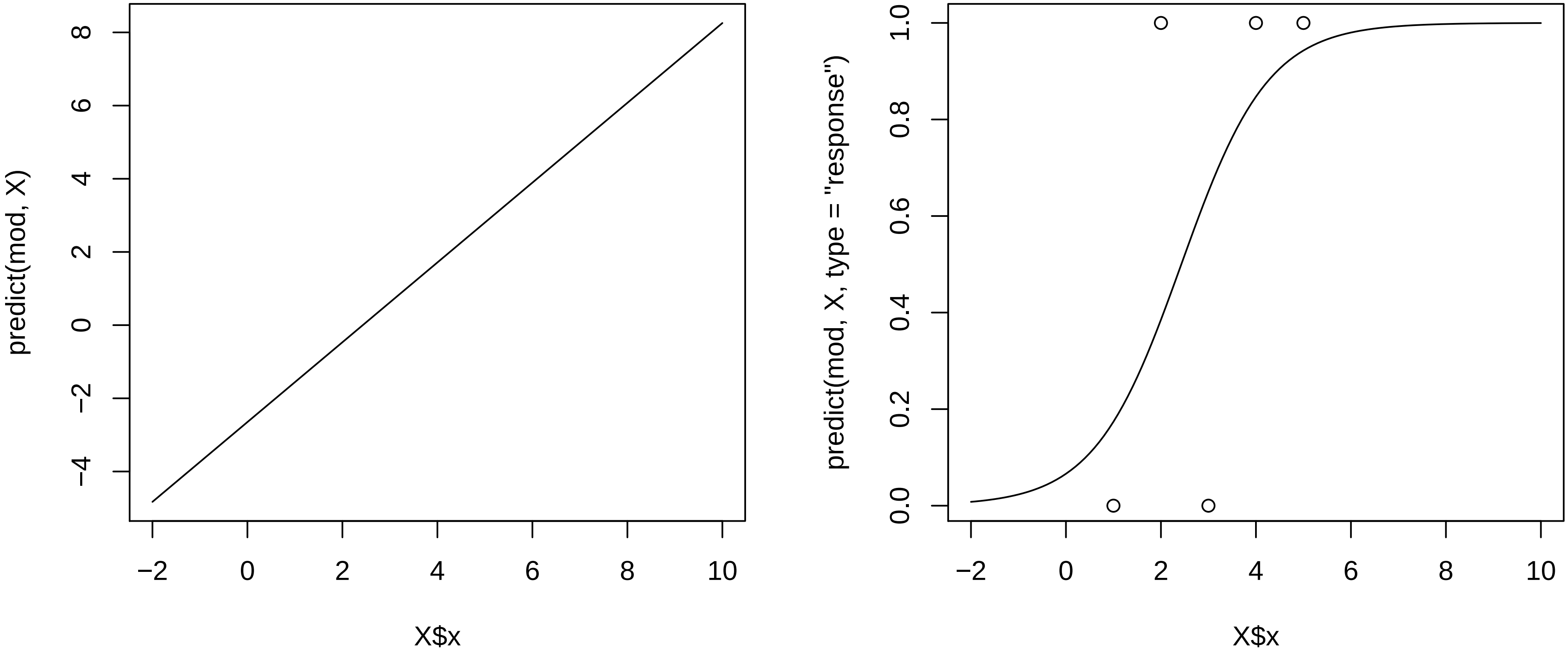
For given values of β, α, and x_i the pdf of the binomial distribution is fully determined, so we can calculate the likelihood.

```
> mod <- glm(y ~ x, family = binomial())
> predict(mod)
         1          2          3          4          5
-1.5581610 -0.4677355  0.6226901  1.7131156  2.8035412
> predict(mod, type = "response")
        1         2         3         4         5
0.1739107 0.3851524 0.6508301 0.8472400 0.9428669
```

```
> mod <- glm(y ~ x, family = binomial())
> predict(mod)
         1          2          3          4          5 
-1.5581610 -0.4677355  0.6226901  1.7131156  2.8035412 
> predict(mod, type = "response")
        1         2         3         4         5 
0.1739107 0.3851524 0.6508301 0.8472400 0.9428669 
```

```
> X <- data.frame(x = seq(0, 10, 0.1))
> plot(X$x, predict(mod, X), type = "l")
> plot(X$x, predict(mod, X, type = "response"), type = "l")
> points(x, y)
```


predict(mod, X)
8
6
4
2
0
−2
−4
−2
0
2
4
6
8
10
X$x


predict(mod, X)
8
6
4
2
0
−2
−4
−2
0
2
4
6
8
10
X$x
predict(mod, X, type = "response")
1.0
0.8
0.6
0.4
0.2
0.0
−2
0
2
4
6
8
10
X$x

Key points to remember:

With maximum likelihood, we have a method to:

- Estimate parameters (the MLEs); these are found either with analytical formulae, or (more often) with numerical algorithms;
- Assess confidence intervals in the MLEs, either with the second derivatives of the likelihood function, or with profile likelihood;
- Compare models with likelihood ratio tests (LRTs) or Akaike information criterion (AIC);
- Apply to wide range of situations as long as there is a random component in the model.