Basic Statistical Analysis in Life and Environmental Sciences

Rodrigo Labouriau

Department of Mathematics, Aarhus University

Module 4, Day 7 - Gaussian Models
2024
(One-way/two-ways classification structures, linear/non-linear regression)

(Bonus: Gamma models)

General Remark

This material is only for internal use in the course.

Please, do not circulate and **do not record**.

Outline

Important concepts:

- Statistical models
- Parameter in statistical model
- Point estimation
- Likelihood function and Maximum likelihood estimate
- Confidence interval and hypothesis test
- Likelihood ratio test
- One-way and two-ways binomial model.
- Binomial models (one-, two-ways, logistic regression and covariance analysis models)
- Poisson models (one-, two-ways and linear/non-linear regression)

Two examples

- We discussed two examples of Poisson models: of similar nature!
- Deaths by horse kick in the Prussian army.
  All the deaths in 20 years (1875-1894)
  Poisson one- and two-ways models.
- Number of colony forming units (CFU) of *Penicillium verrucosum* in soil.

  (Elmholt, Labouriau, Hestbjerg and Jørgensen, 1998).

  Poisson regression models (linear and quadratic)

- A random variable Y is said to follow a *Poisson distribution* with parameter λ ($\lambda > 0$) if

$$P(Y = y) = \frac{e^{-\lambda}\lambda^y}{y!},$$

for $y = 0, 1, 2, \ldots$.
Here $y! = y \cdot (y-1) \cdot \cdots \cdot 1$ and $0! = 1$.

- A Poisson variable takes only non-negative integer values.
The Poisson distribution describes typically counts
(but there exist many other distributions for counts!)

- Notation: $Y \sim Po(\lambda)$

- $E(Y) = Var(Y) = \lambda$

Poisson as a law of rare events

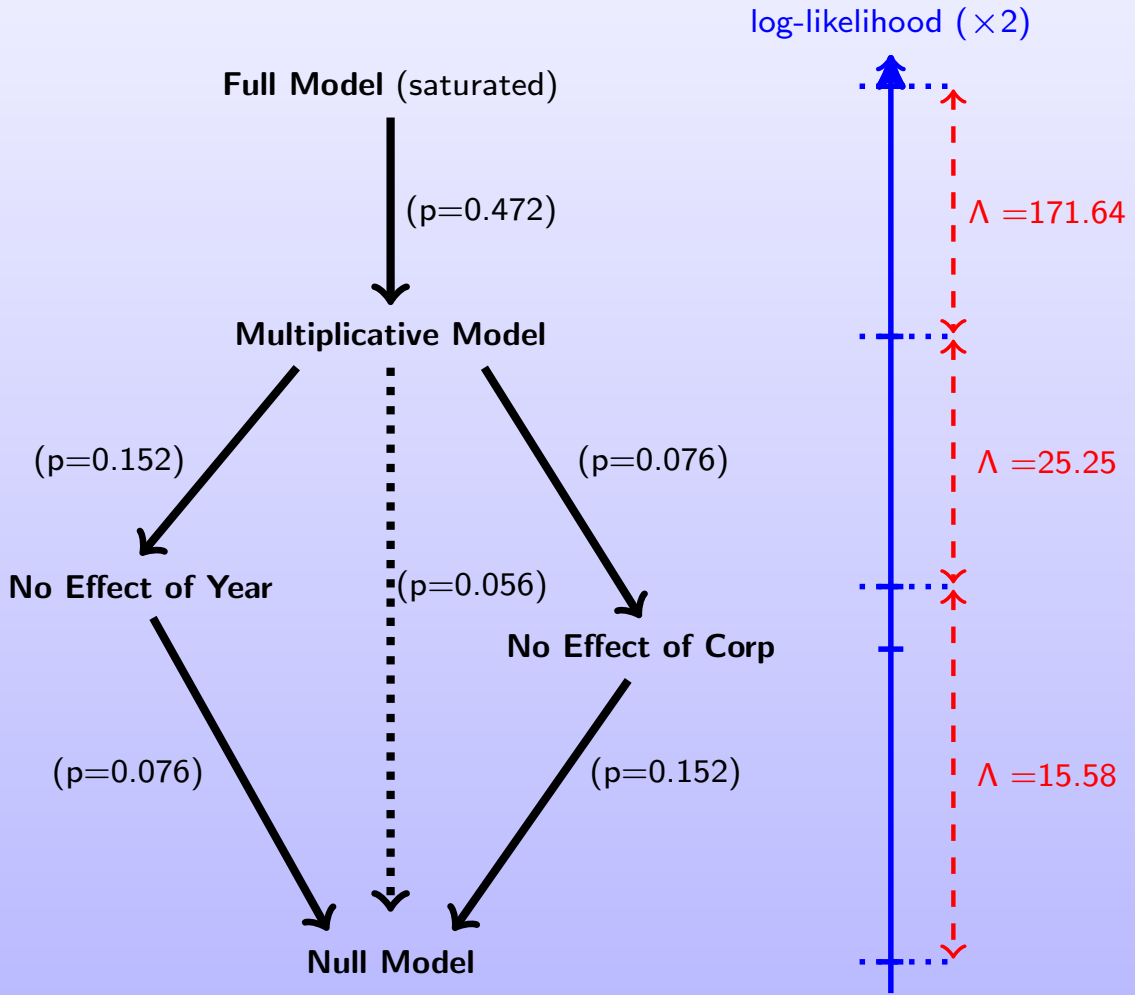
- Suppose that we observe a binomial random variable, $Y \sim Bi(n, p)$.
- Suppose that $n \to \infty$ and $p \to 0$ in such a way that np remains finite and tends to a number λ (i.e., $np \to \lambda$), then the probability law of Y tends to a Poisson distribution.
- Example: The number of deaths by horse kicks.

Deaths					
0	1	2	3	4	≥ 5
109	65	22	3	1	0

Rare events!

(122 occurrences in 20 years 6.1 / year 0.61 per corp year)

Poisson two ways classification model



Example: Penicillium in soil

Example: Penicillium in soil

- We performed the following experiment:
  - Make a suspension of the soil;
  - Take successive dilutions of the suspension;
  - Plate the dilutions in Petri dishes and count the number of colonies that appeared after an incubation time.
- This technique is called the plating method (Fisher, 1922).
- Knowing the amount of soil added, estimate the number of CFU / g soil
- The probability distribution of the number of colonies per Petri dish can be deduced (under some reasonable assumptions)!

Poisson deduced from simple assumptions

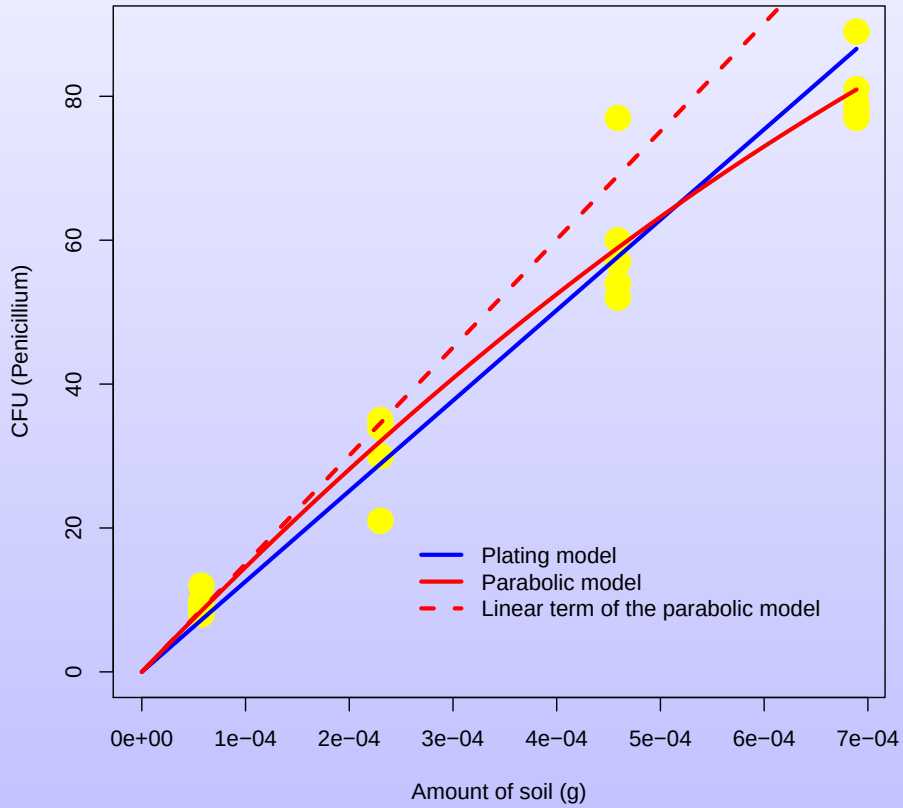
- The probability distribution of the number of colonies per Petri dish can be deduced (under some reasonable assumptions)!
- The number of CFUs in a portion of the suspension is a random quantity denoted Y.
- We assume that:
  - Homogeneous distribution of the CFUs in the suspension.
  - The number of CFUs in two disjoint portions of the suspension are independent
  - The CFUs are not clustered together.
- Under these assumptions it can be shown that the number of CFUs in the Petri dish is distributed according to a Poisson distribution.

(formal proof: differential equations and some basic stochastic processes)

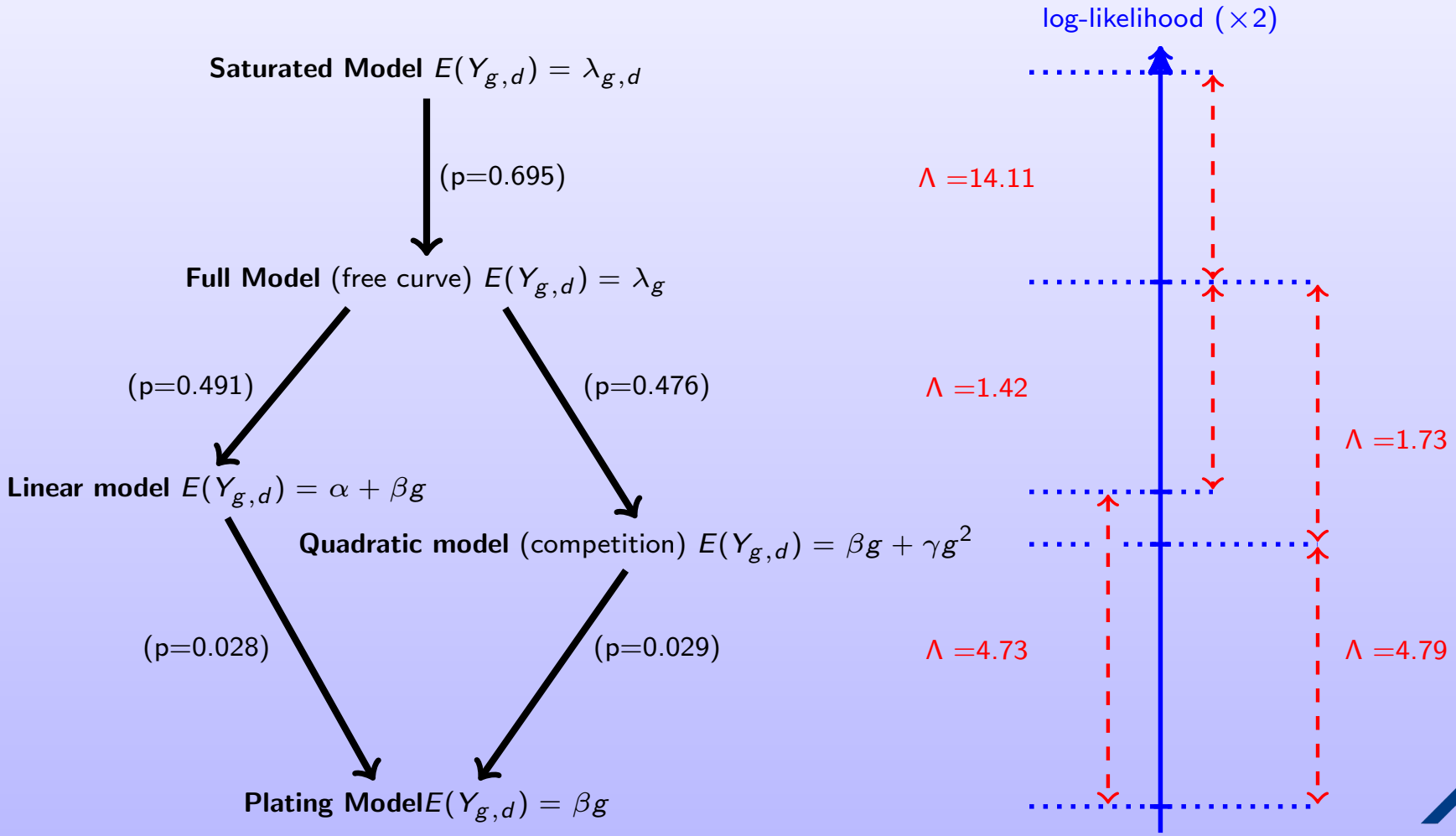
Example: Penicillium in soil

- $Y_{g,d}$ represents the number of Penicillium CFU observed in the dth Petry dish, for which it was added g grams of soil.
- $Y_{g,d} \sim$ Poisson
- Platting method model (linear):
  $Y_{g,d} \sim Po(\lambda_{g,d})$
  $E(Y_{g,d}) = \lambda_{g,d} = \beta g$

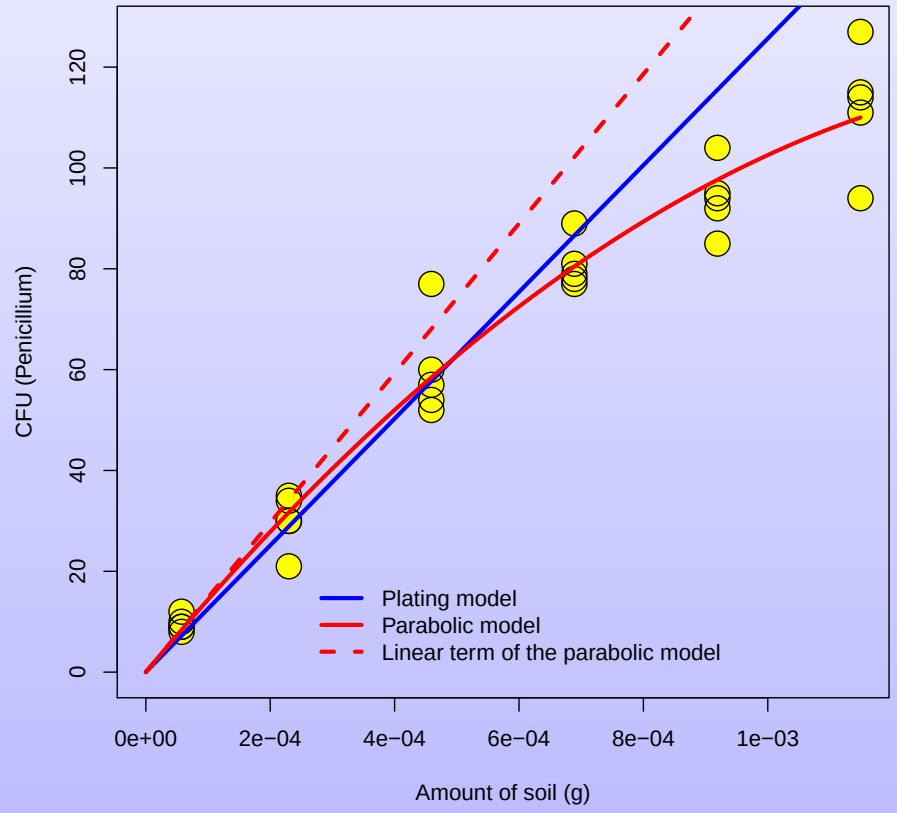
  Interpretation of β: Number of CFU per gram soil! (why?)
- Platting method model with competition/inhibition (quadratic):
  $Y_{g,d} \sim Po(\lambda_{g,d})$
  $E(Y_{g,d}) = \lambda_{g,d} = \beta g + \gamma g^2$



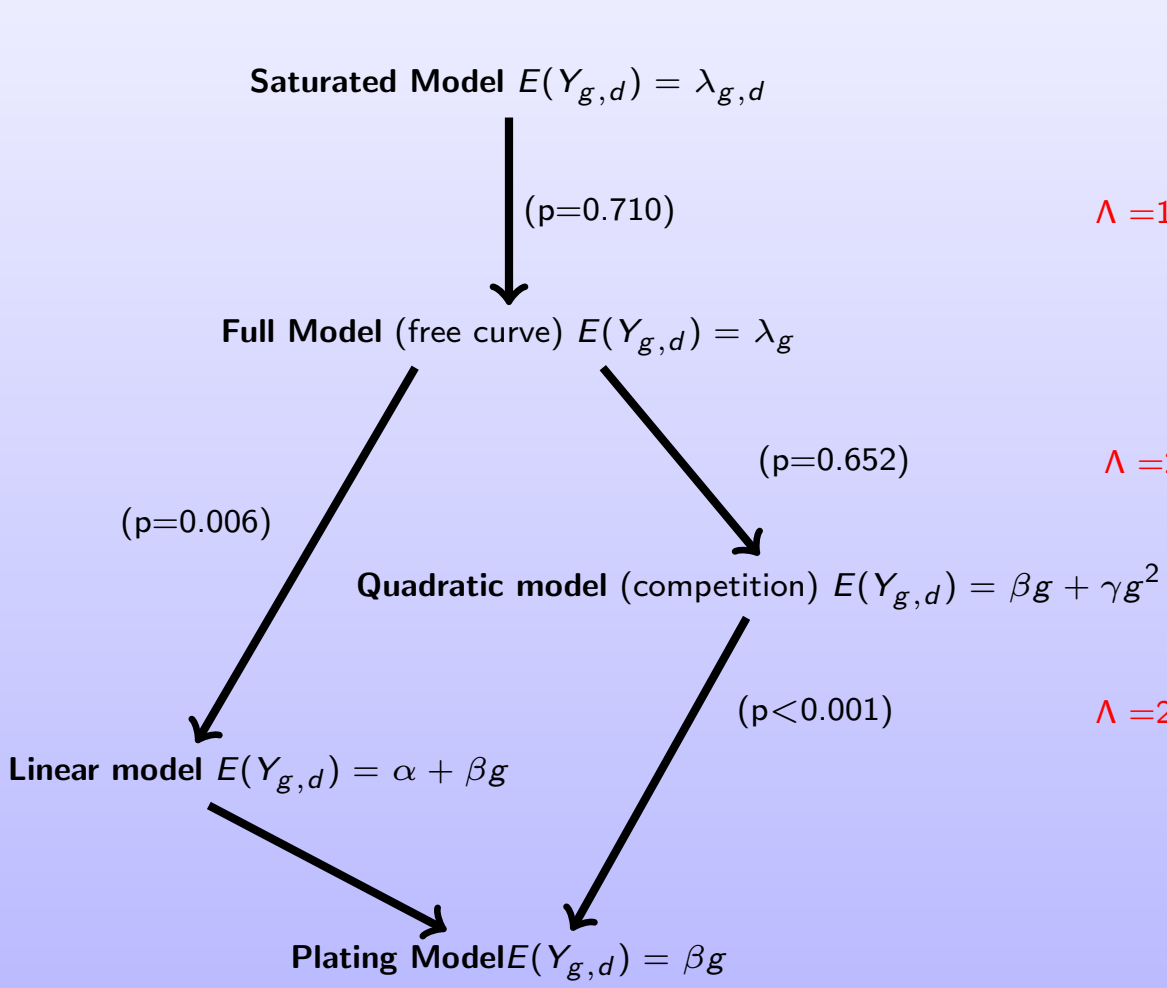
Example: Penicillium in soil

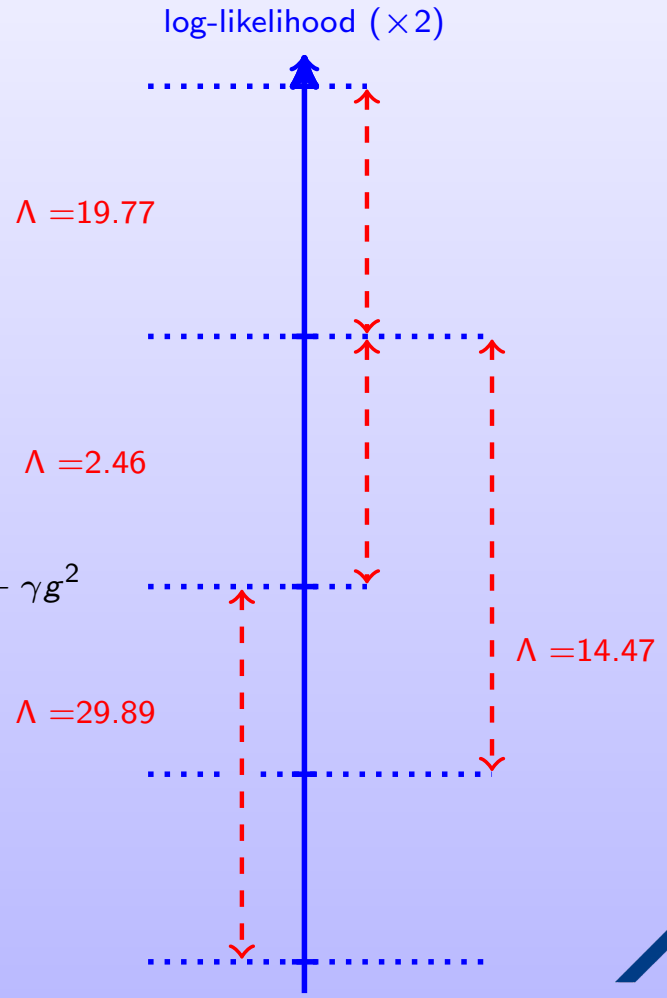


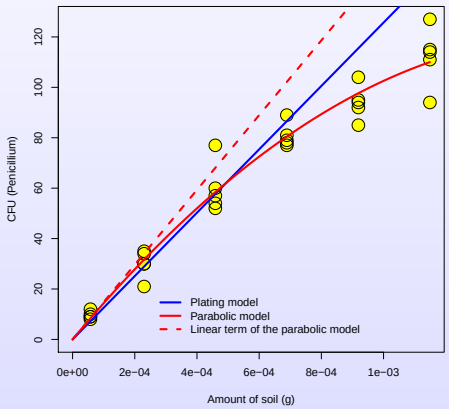
Extended experiment

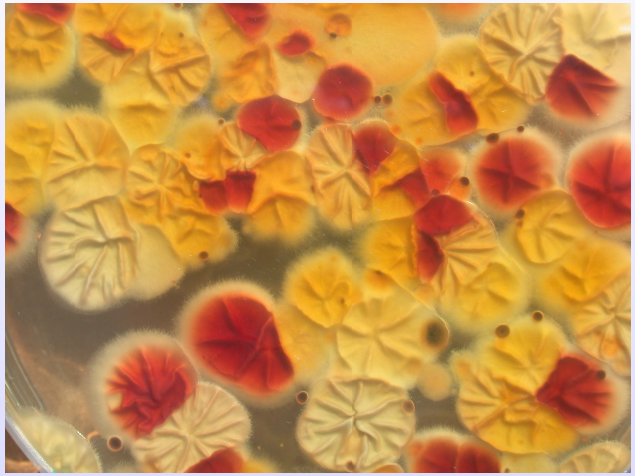


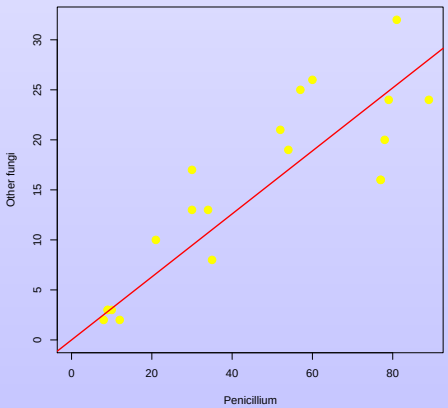
Example: Penicillium in soil - Extended experiment











In conclusion,
Penicillium verrucosum is not like *Homo sapiens sapiens*,
when there is lack of resources they do not kill the other species!

Normal distribution

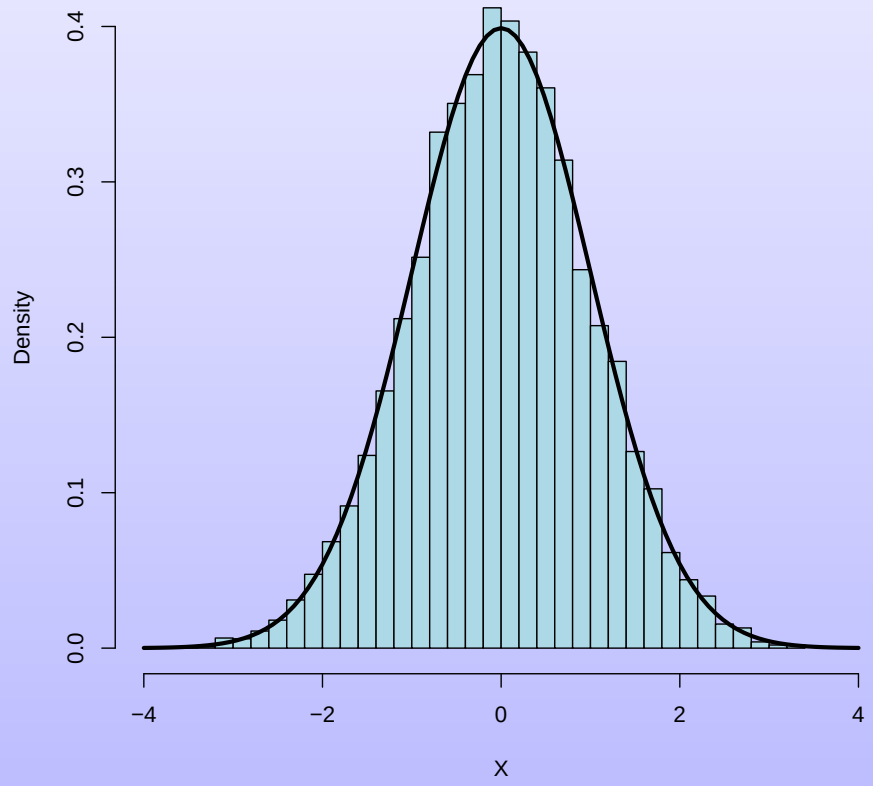
- Central distribution among the continuous distributions
  Two reasons:
  - Central Limit Theorem:
    Approximate well many cases
  - Easy to compute:
    Maximum likelihood estimate is the mean or least squares estimates

    Calculations can be done with pocket calculator and a couple of tables
- Normal distribution: continuous distribution depending on two parameters, μ and σ^2 and probability density given by, for each real number x,

$$\phi(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}.$$

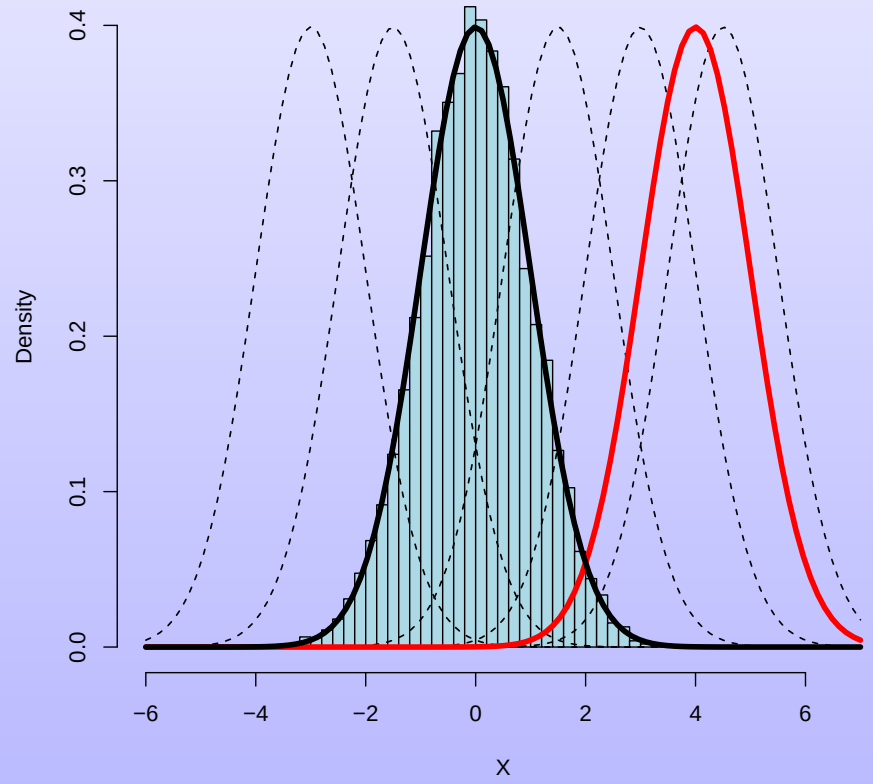
  Here μ is a real number and σ is a positive number ($\sigma > 0$).
- $X \sim N(\mu,\sigma^2)$ $\quad E(X) = \mu,\ Var(X) = \sigma^2$
  (the variance is not a function of the mean).

The normal distribution - Simulated superposed with the density



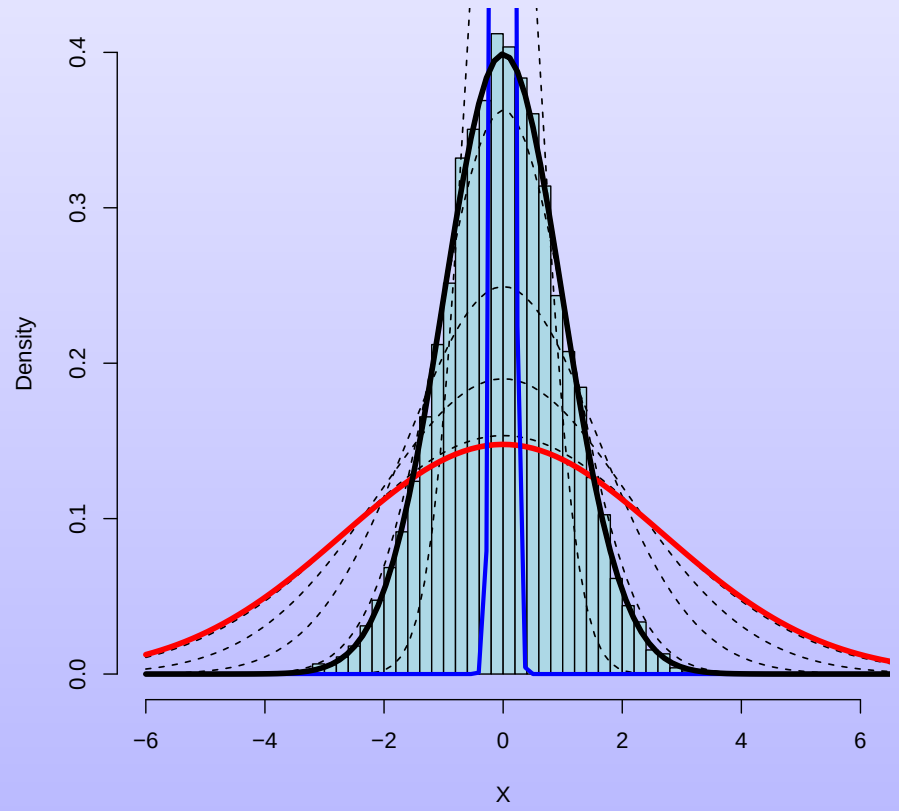
The normal distribution

Simulated samplesuperposed with the density with different means



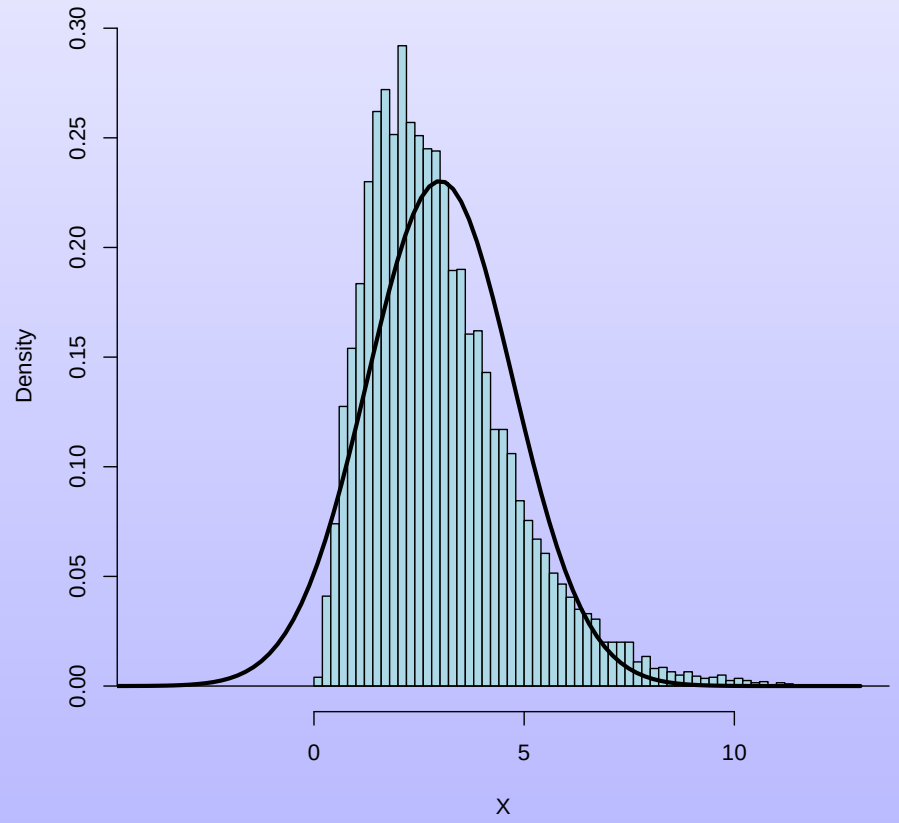
The normal distribution

Simulated sample superposed with the density with different variances



The normal distribution

Simulated gamma superposed with the normal density with the same mean and variance



Normal distribution - Notation

$$X \sim N(\mu, \sigma^2)$$

$$E(X) = \mu,\ Var(X) = \sigma^2$$

Suggestion: Run the tutorial "Stat-Tutorial-04" for getting intuition on the normal distribution

The central limit theorem

- Consider $X_1, X_2, \ldots$ are independent and identically distributed random variables for which $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$, where $0 < \sigma^2 < \infty$

- The central limit theorem states that (under the assumptions above), for n sufficiently large

$$X_1 + \cdots + X_n$$

is approximately normally distributed.

Or equivalently,

$$\frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

follows approximately a standard normal distribution, a normal distribution with mean 0 and variance 1

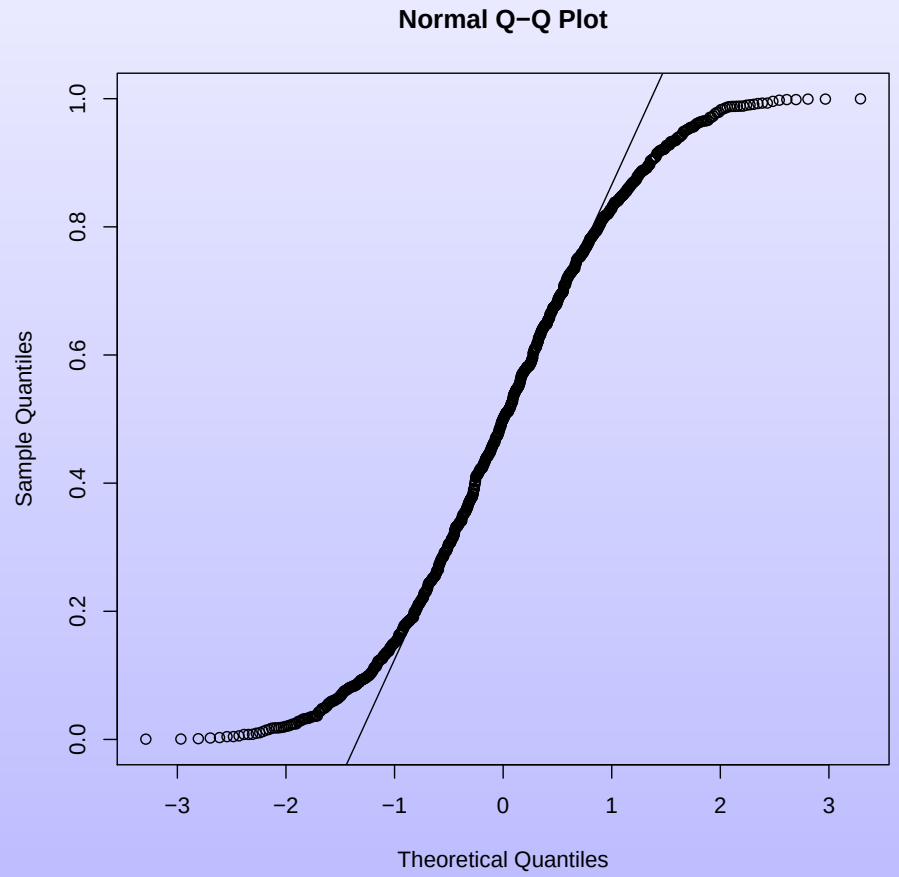
Simulating observations from a uniformly distributed

```
n.observations <- 500
```

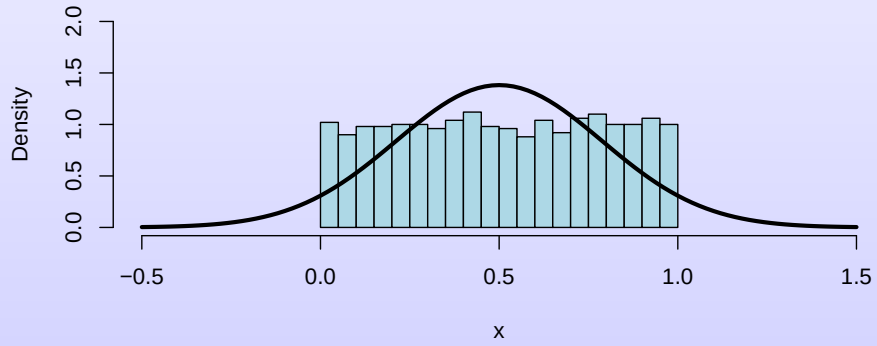
```
x <- runif(n.observations)
```

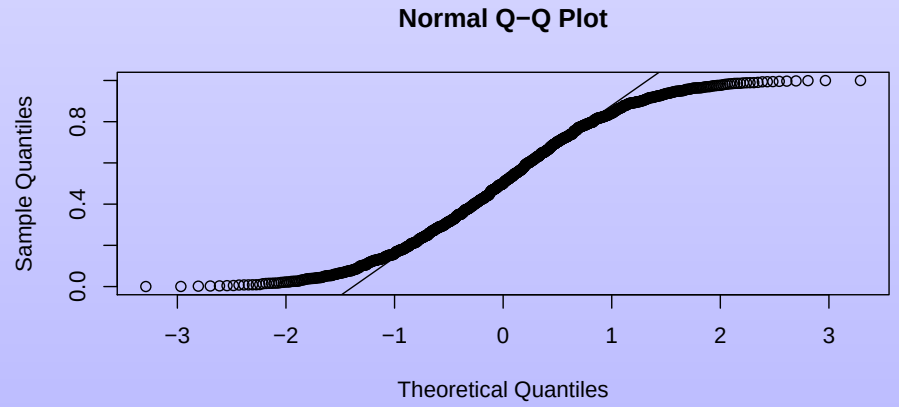
```
qqnorm(x); qqline(x)
```

Normal QQ-plot of simulated uniform distribution



Normal QQ-plot of simulated uniform distribution



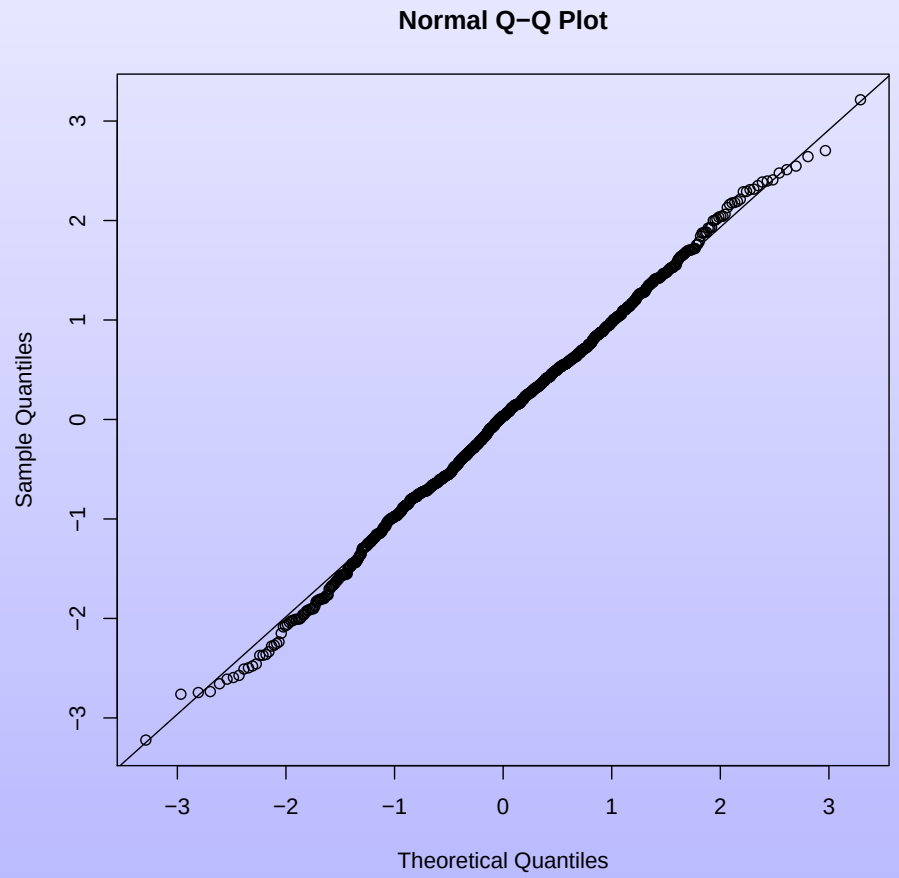


Normal distribution

Central Limit theorem for uniform distributed variables

```
# Remark: Note that the expectation of a uniform distributed random variable
#         is 1/2 = 0.5 and the variance is 1/12 (I did not make this calculation
#         in the lectures, but believe me, please)

n.rep <- 1000                                    # Number of replicates
X <- numeric(n.rep)
n.observations <- 50                             # Number observations
for(i in 1:n.rep){
  x <- runif(n.observations)
  X[i] <- (sqrt(n.observations)*(mean(x) - 0.5)) / sqrt(1/12)
}
qqnorm(X); qqline(X)
```

Normal QQ-plot of the **means** of simulated uniformly distributed random variables

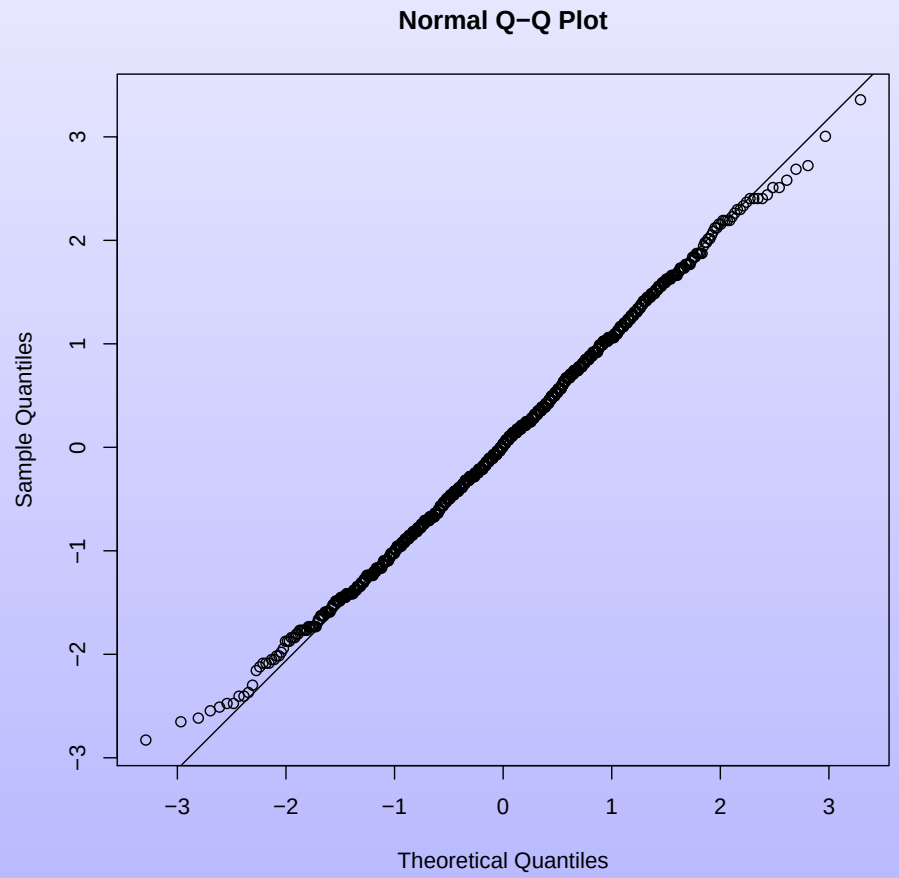


Normal distribution

Central Limit theorem for Poisson distributed variables

```
n.rep <- 1000
X <- numeric(n.rep)
L <- 4              # This will be the intensity or lambda parameter.
n.observations <- 200
for(i in 1:n.rep){
   x <- rpois(n=n.observations, lambda=L)
   X[i] <- (sqrt(n.observations)*(mean(x) - L)) / sqrt(L)
}
```

Normal QQ-plot of the **means** of simulated Poisson distributed random variables

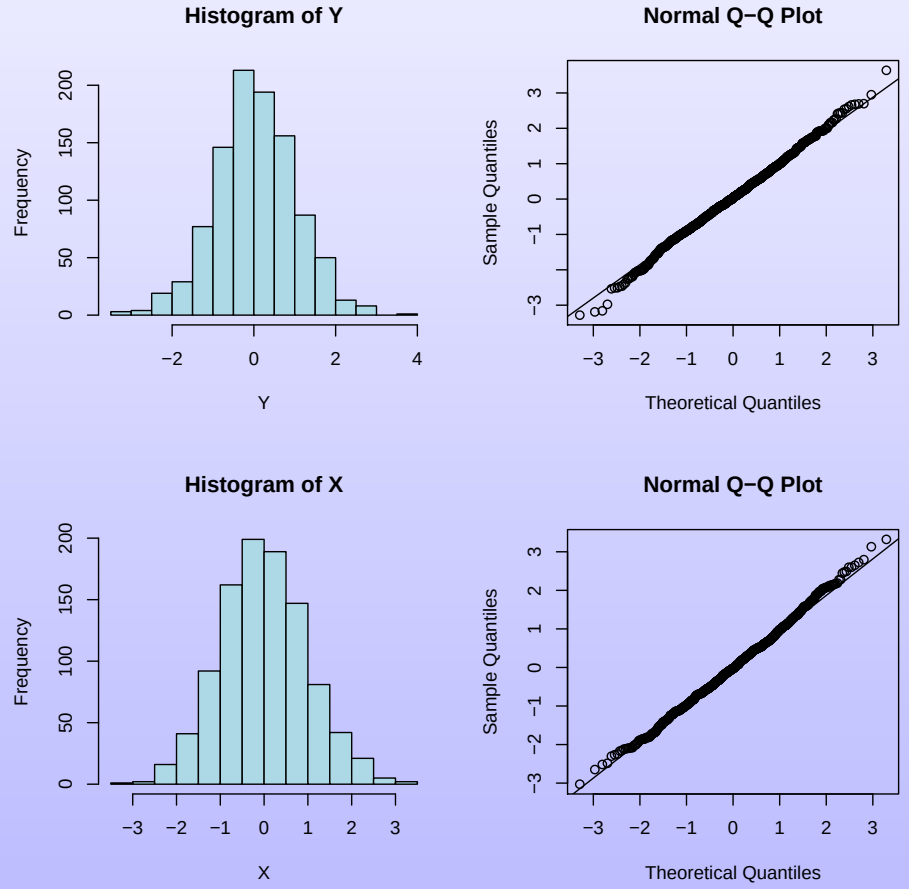


Normal distribution

Central Limit theorem for Cauchy distributed variables

```
Y <- rnorm(1000);X <- rnorm(1000)
par(mfrow=c(2,2))
hist(Y, col = "lightblue")
qqnorm(Y);qqline(Y)
hist(X, col = "lightblue")
qqnorm(X);qqline(X)
```

Normal QQ-plot of two simulated normally distributed r.v.



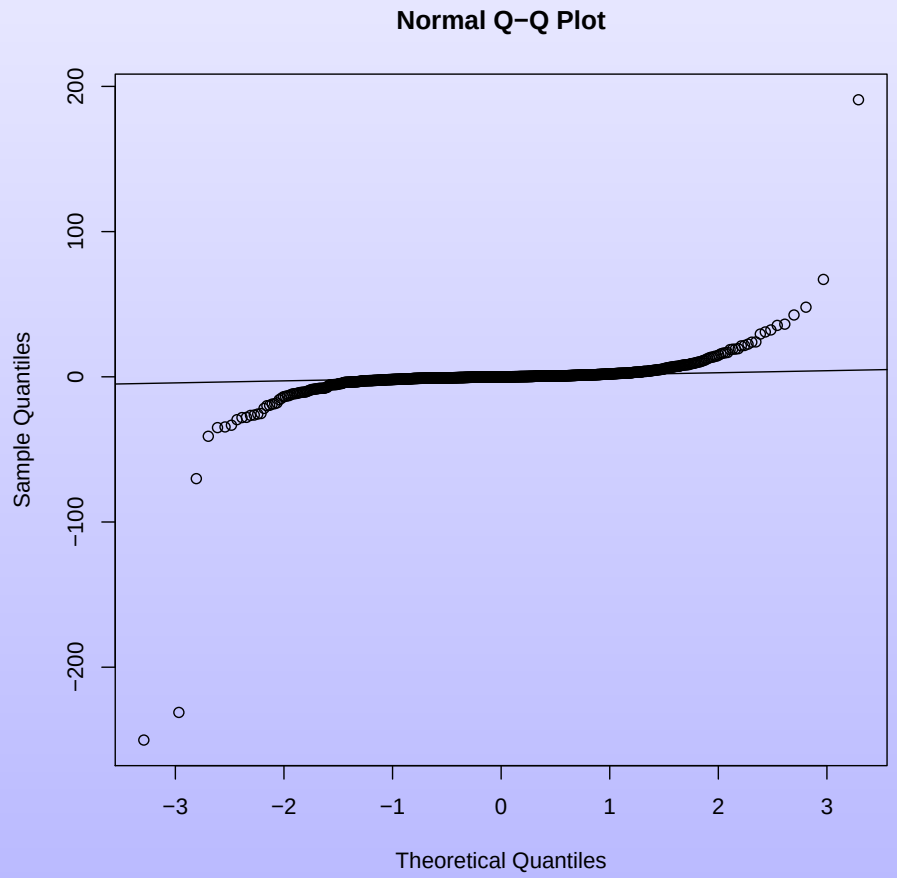
Normal distribution

The ratio of two normal distributed r.v. is **not** normally distributed

```
Z <- Y/X
```

```
qqnorm(Z);qqline(Z)
```

The ratio of two normal distributed r.v. is **not** normally distributed, but Cauchy distributed

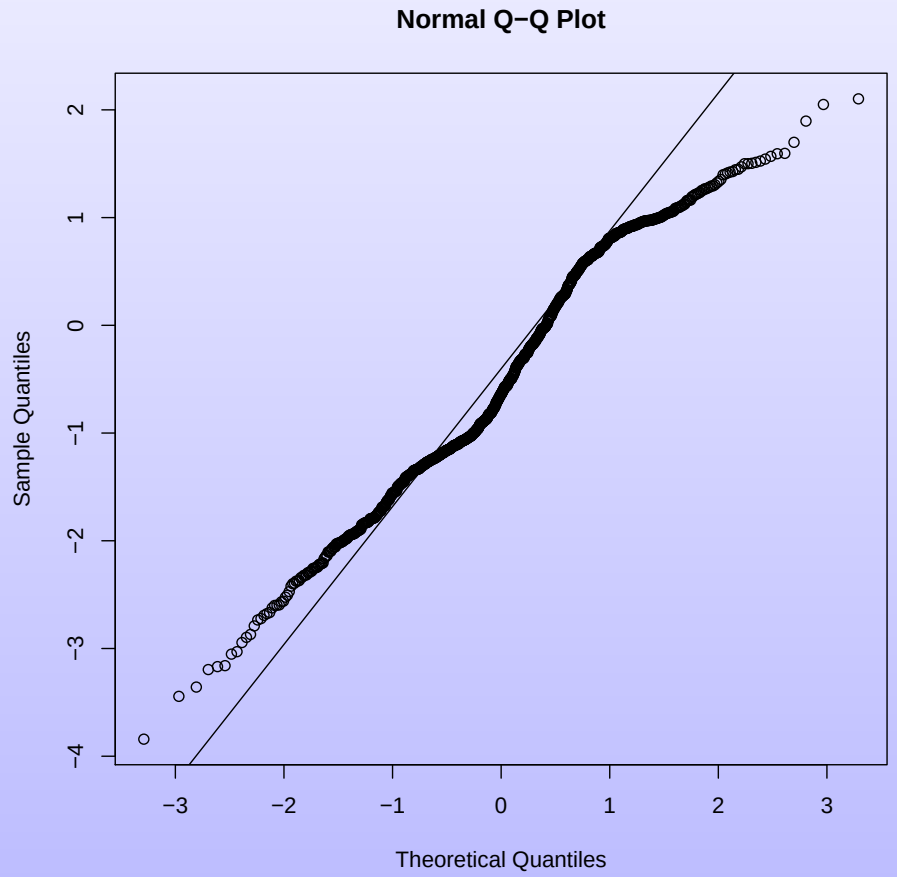


Normal distribution

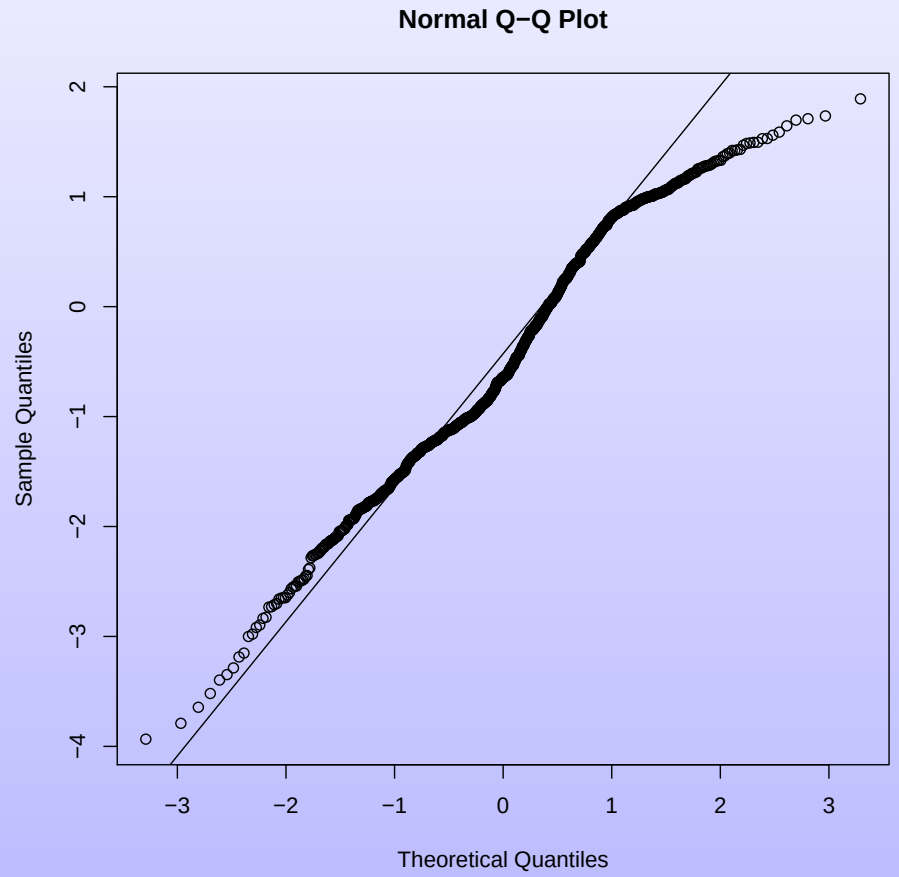
Trying to use the Central Limit theorem for means of Cauchy distributed variables

```
n.rep <- 1000
Z <- numeric(n.rep)
n.observations <- 1000
for(i in 1:n.rep){
   y <- rnorm(n.observations)
   x <- rnorm(n.observations)
   z <- y/x
   Z[i] <- (sqrt(n.observations)*(mean(z) - 0.5)) / sqrt(var(z))
}
qqnorm(Z); qqline(Z)
```

Normal QQ-plot of simulated **means** of Cauchy distributed r.v.



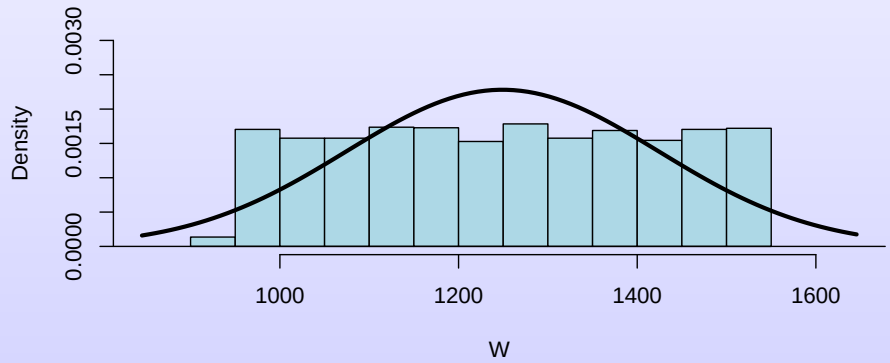
Normal QQ-plot of simulated **means** of Cauchy distributed r.v. (100,000 repetitions!)

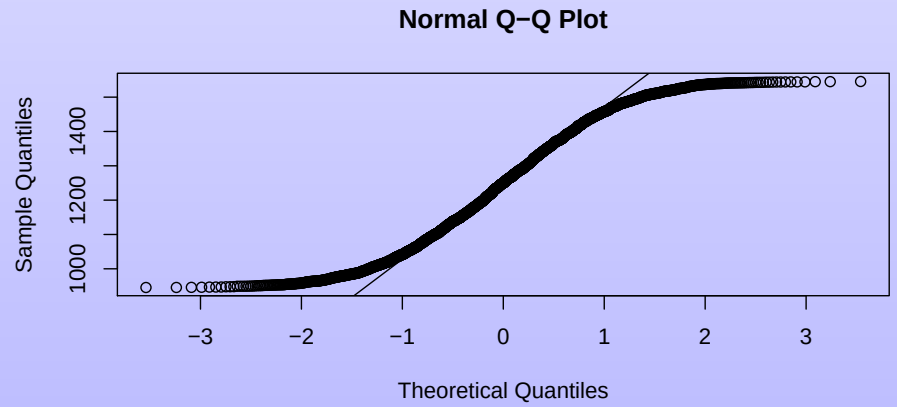


One-way ANOVA the distribution of the individual weights

- Weights of *Dolichos biflorus* seeds a leguminosae
  (selected for uniformity)
- Automatic weighting of seeds
- 50 batches of 50 seeds each
  Recorded the weight of each of the 2,500 seeds
  1-2 g per seed (measured in mg)

One-way ANOVA the distribution of the individual weights





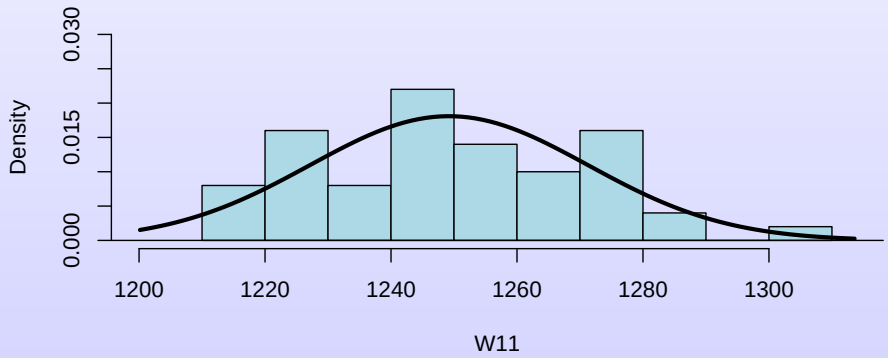
(P-value of Shapiro-Wilks test smaller than $2.2\,10^{-16}$)

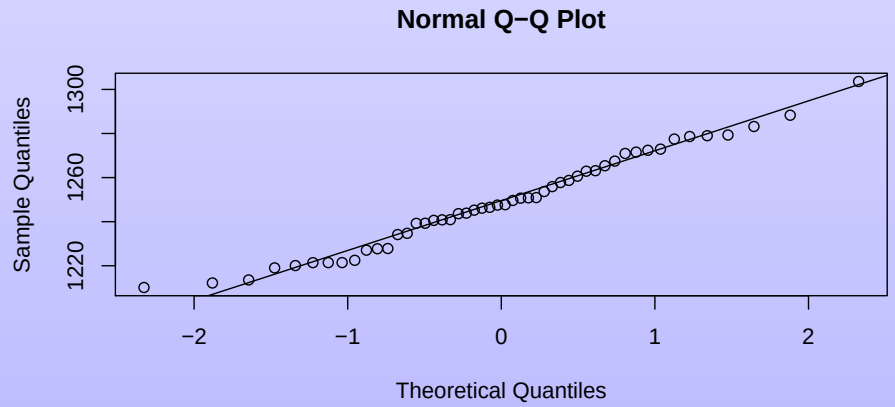
One-way ANOVA the distribution of the batchee's weights

- The distribution of the individual seed weights is clearly NOT normally distributed
- 50 batches of 50 seeds each
- Due to the Central Limit Theorem, taking averages per batch we might expect to obtain approximately normally distributed results

(averaging is equivalent to summing and rescaling)

One-way ANOVA the distribution of the batch averaged weights





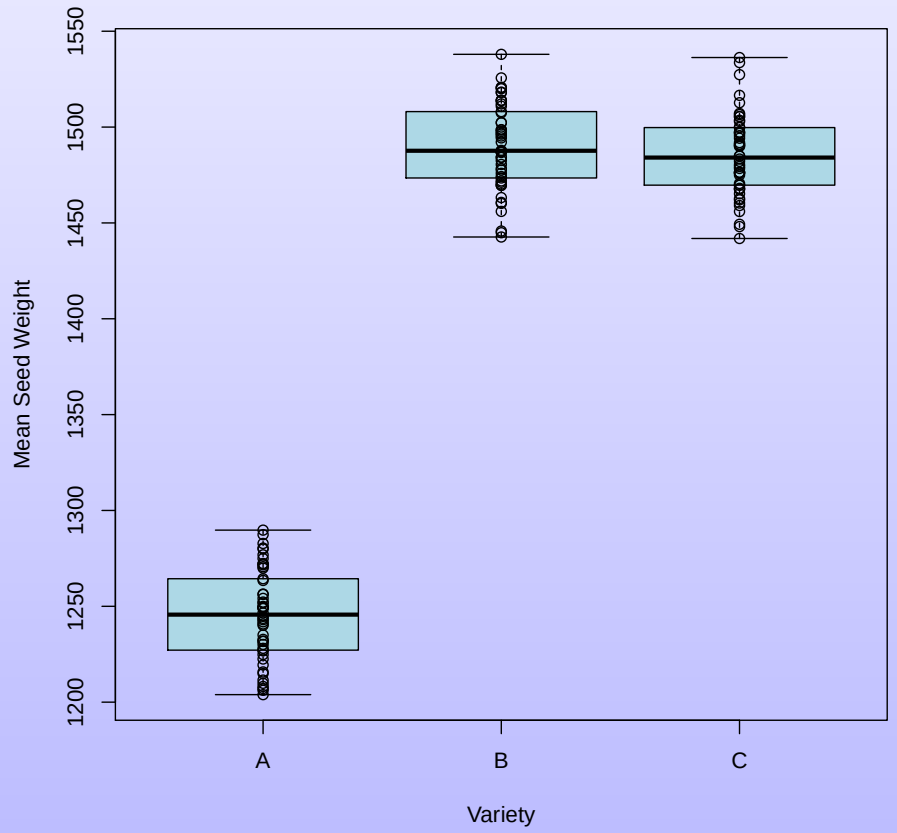
(P-value of Shapiro-Wilks test = 0.5372)

One-way ANOVA - comparing three varieties

- The data of this example is more complex!
- There are 150 batches and three varieties of *Dolichos biflorus* 50 batches of each varieties
- A balanced design

```
> str(DolichosOneWay)
'data.frame':        150 obs. of  2 variables:
 $ Y      : num  1264 1487 1534 1275 1521 ...
 $ Variety: Factor w/ 3 levels "A","B","C": 1 2 3 1 2 3 1 2 3 1 ...
```

One-way ANOVA - comparing three varieties



One-way ANOVA - comparing three varieties

- Y_{vb} is the random variable representing the averaged weight of the b^{th} batch ($b = 1, ..., 50$) of the v^{th} variety ($v = A, B, C$)
- The model assumes that the random variables $Y_{A1}, \ldots, Y_{C50}$ are:
  - independent,
  - normally distributed
  - have the same variance (say $Var(Y_{vb}) = \sigma^2$)
  - have expectation depending only on the variety (say $E(Y_{vb}) = \tau_v$)
- In short,

$$Y_{vb} \sim N\left(\tau_v, \sigma^2\right), \text{ for } v = A, B, C \text{ and } b = 1, ..., 50,$$

where $Y_{A1}, \ldots, Y_{C50}$ are independent.

One-way ANOVA - comparing three varieties

```
> M <- glm(Y ~ Variety + 0, family = gaussian(link = "identity"), data = D)

> summary(M)

Call: glm(formula = Y ~ Variety + 0, family = gaussian(link = "identity"),  data = D)

...

Coefficients:

         Estimate Std. Error t value Pr(>|t|)

VarietyA 1245.258      3.195   389.7   <2e-16 ***

VarietyB 1489.787      3.195   466.3   <2e-16 ***

VarietyC 1485.278      3.195   464.9   <2e-16 ***

---

...
```

One-way ANOVA - comparing three varieties - some model control

```
> M <- glm(Y ~ Variety + 0, family = gaussian(link = "identity"), data = D)
```

```
> Residuals <- residuals(M, "response")
```

```
> Fitted    <- fitted(M)
```

```
> library(car)
```

```
> qqPlot(Residuals)
```

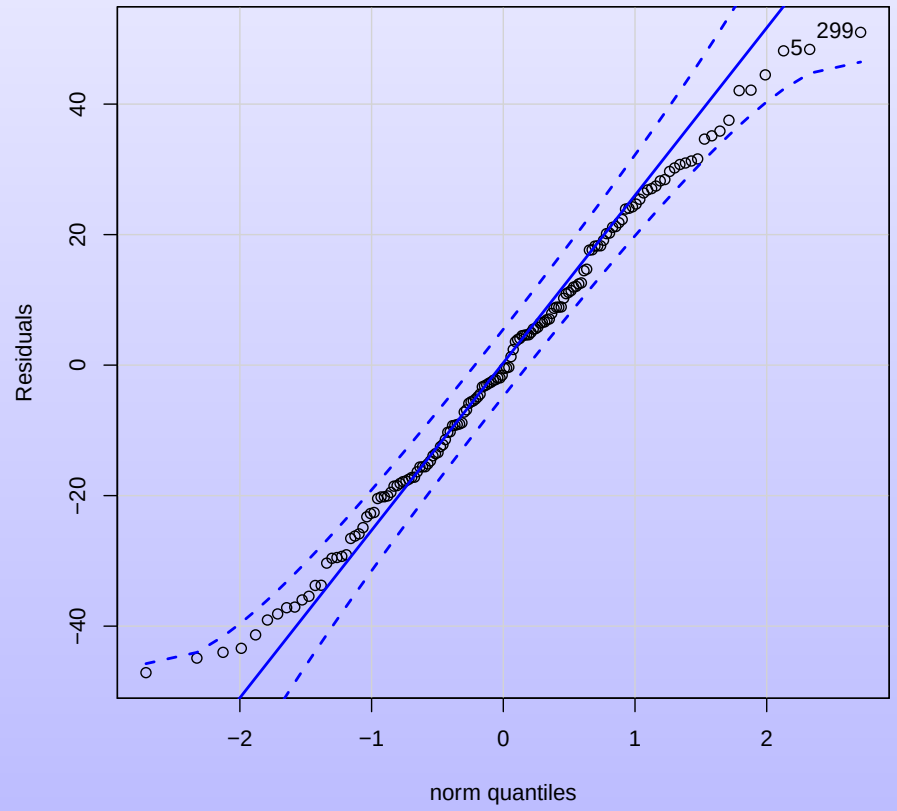
```
> shapiro.test(Residuals)
```

```
        Shapiro-Wilk normality test
```

```
data:  Residuals
```

```
W = 0.98919, p-value = 0.3016
```

One-way ANOVA - comparing three varieties - some model control



One-way ANOVA - comparing three varieties - some model control

```
> M <- glm(Y ~ Variety + 0, family = gaussian(link = "identity"), data = D)
```

```
> Residuals <- residuals(M, "response")
```

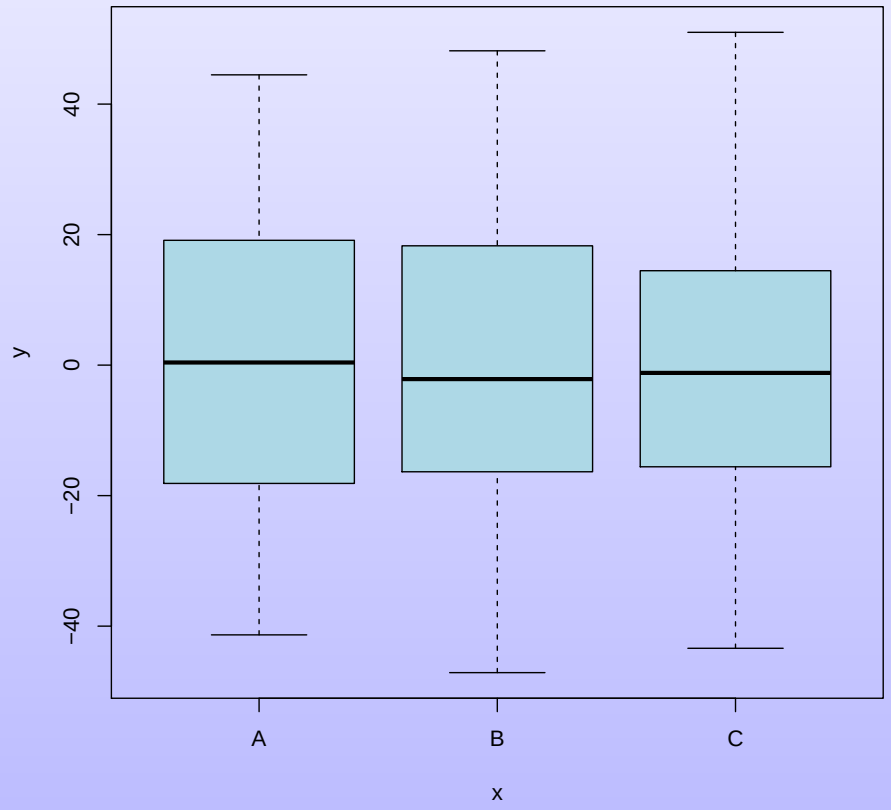
```
> Fitted    <- fitted(M)
```

```
> plot(D$Variety, Residuals, col = "lightblue")
```

```
> bartlett.test(Residuals, g = D$Variety)
```

```
        Bartlett test of homogeneity of variances

data:  Residuals and D$Variety

Bartlett's K-squared = 0.57806, df = 2, p-value = 0.749
```

One-way ANOVA - comparing three varieties - some model control



One-way ANOVA - the null model

- Y_{vb} is the random variable representing the averaged weight of the b^{th} batch ($b = 1, ..., 50$) of the v^{th} variety ($v = A, B, C$)
- The model assumes that the random variables $Y_{A1}, \ldots, Y_{C50}$ are:
  - independent,
  - normally distributed
  - have the same variance (say $Var(Y_{vb}) = \sigma^2$)
  - have the same expectation (say $E(Y_{vb}) = \tau$)
- In short,

$$Y_{vb} \sim N\left(\tau, \sigma^2\right), \text{ for } v = A, B, C \text{ and } b = 1, ..., 50,$$

where $Y_{A1}, \ldots, Y_{C50}$ are independent.

One-way ANOVA - testing for possible differences between varieties

- Idea: test the (possible) differences between the varieties by comparing the two models below
- One-way analysis of variance model: $Y_{vb} \sim N\left(\tau_v, \sigma^2\right)$
  Null model: $Y_{vb} \sim N\left(\tau, \sigma^2\right)$

```
> anova(M0, M, test = "F")
```

```
Analysis of Deviance Table
```

```
Model 1: Y ~ 1
```

```
Model 2: Y ~ Variety + 0
```

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
1	149	2032110			
2	147	75031	2	1957079	1917.1 < 2.2e-16 ***

One-way ANOVA - post-hoc analysis

```
> TT <- posthoc(M, EffectLabels = levels(D$Variety))
```

```
> print(TT)
```

```
  Levels                       ParameterCI
1      A 1245.2577(1238.9955-1251.5199)a
2      B 1489.7867(1483.5246-1496.0489)b
3      C 1485.2784(1479.0162-1491.5406)b
```

Two-ways ANOVA - comparing three varieties in two fields

- The data analysed above is only partial!
- There are 300 batches and three varieties of *Dolichos biflorus* in two fields

  50 batches of each varieties in each field

```
> str(DolichosTwoWays)
'data.frame':        300 obs. of  3 variables:
 $ Y      : num  1264 1489 1487 1746 1534 ...
 $ Variety: Factor w/ 3 levels "A","B","C": 1 1 2 2 3 3 1 1 2 2 ...
 $ Field  : Factor w/ 2 levels "I","II": 1 2 1 2 1 2 1 2 1 2 ...
```

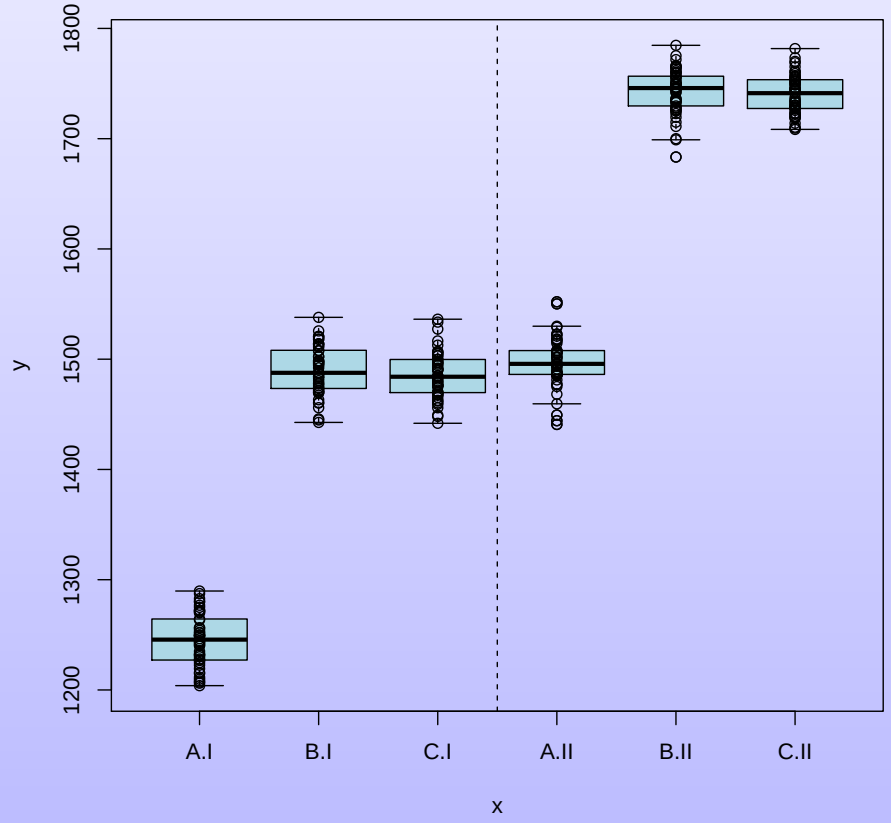
Two-ways ANOVA - comparing three varieties in two fields

```
> D <- DolichosTwoWays
```

```
> table(D$Variety, D$Field)
```

```
   I II
A 50 50
B 50 50
C 50 50
```

Two-ways ANOVA - comparing three varieties in two fields



Two-ways ANOVA - comparing three varieties in two fields - Interaction Model

- Y_{vfb} is the random variable representing the averaged weight of the b^{th} batch $(b = 1, ..., 50)$ of the v^{th} variety $(v = A, B, C)$ from the f^{th} field $(f = I, II)$
- The model assumes that the random variables $Y_{AI1}, \ldots, Y_{CII50}$ are:
  - independent,
  - normally distributed
  - have the same variance (say $Var(Y_{vfb}) = \sigma^2$)
  - have expectation depending on the combination of variety and field (say $E(Y_{vfb}) = \tau_{vf}$)
- In short,

$$Y_{vfb} \sim N\left(\tau_{vf}, \sigma^2\right), \text{ for } v = A, B, C, \ f = I, II \text{ and } b = 1, ..., 50,$$

where $Y_{AI1}, \ldots, Y_{CII50}$ are independent.

Two-ways ANOVA - comparing three varieties in two fields - Interaction Model

```
> Minter <- glm(Y ~ Variety + Field + Variety:Field,  family = gaussian(link = "identity"), data = D)
```

```
> Minter1 <- glm(Y ~ Variety * Field, data = D)
```

```
> Minter2 <- glm(Y ~ Variety + Field + Variety:Field + 0 , data = D)
```

```
> Minter3 <- glm(Y ~ Variety:Field + 0 , data = D)
```

```
> deviance(Minter);deviance(Minter1);deviance(Minter1);deviance(Minter3)
```

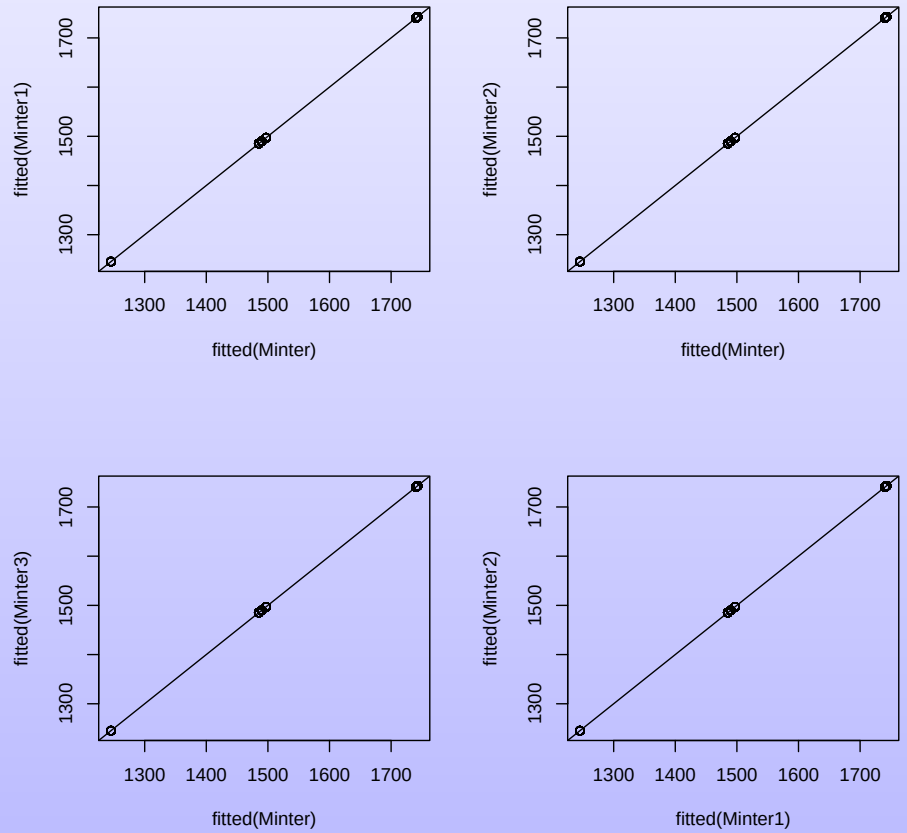
```
[1] 136720.7
```

```
[1] 136720.7
```

```
[1] 136720.7
```

```
[1] 136720.7
```

Two-ways ANOVA - comparing three varieties in two fields - Interaction Model



Two-ways ANOVA - comparing three varieties in two fields - Interaction Model

```
> summary(Minter3)

Call: glm(formula = Y ~ Variety:Field + 0, data = D)

Coefficients:

                Estimate Std. Error t value Pr(>|t|)
VarietyA:FieldI   1245.26       3.05   408.3   <2e-16 ***
VarietyB:FieldI   1489.79       3.05   488.5   <2e-16 ***
VarietyC:FieldI   1485.28       3.05   487.0   <2e-16 ***
VarietyA:FieldII  1497.03       3.05   490.9   <2e-16 ***
VarietyB:FieldII  1742.87       3.05   571.5   <2e-16 ***
VarietyC:FieldII  1740.40       3.05   570.7   <2e-16 ***
```

Two-ways ANOVA - comparing three varieties in two fields - Interaction Model

```
> summary(Minter2)

Call: glm(formula = Y ~ Variety + Field + Variety:Field + 0, data = D)

Coefficients:

                 Estimate Std. Error t value Pr(>|t|)
VarietyA         1245.258      3.050 408.320   <2e-16 ***
VarietyB         1489.787      3.050 488.501   <2e-16 ***
VarietyC         1485.278      3.050 487.023   <2e-16 ***
FieldII           251.774      4.313  58.377   <2e-16 ***
VarietyB:FieldII    1.308      6.099   0.215    0.830
VarietyC:FieldII    3.345      6.099   0.548    0.584
```

Two-ways ANOVA - comparing three varieties in two fields - Interaction Model

```
> summary(Minter1)

Call: glm(formula = Y ~ Variety * Field, data = D)

Coefficients:

                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1245.258      3.050 408.320   <2e-16 ***
VarietyB         244.529      4.313  56.697   <2e-16 ***
VarietyC         240.021      4.313  55.651   <2e-16 ***
FieldII          251.774      4.313  58.377   <2e-16 ***
VarietyB:FieldII   1.308      6.099   0.215    0.830
VarietyC:FieldII   3.345      6.099   0.548    0.584
```

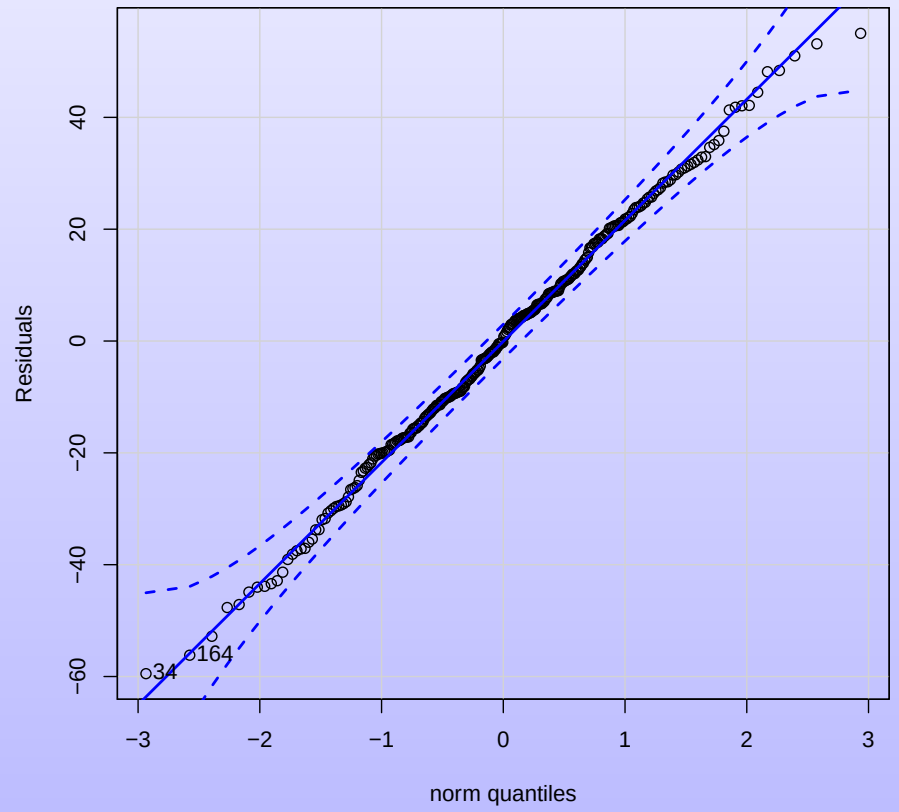
Two-ways ANOVA - comparing three varieties in two fields - Interaction Model

```
> summary(Minter)

Call: glm(formula = Y ~ Variety + Field + Variety:Field, data = D)

Coefficients:

                Estimate Std. Error t value Pr(>|t|)
(Intercept)     1245.258      3.050 408.320   <2e-16 ***
VarietyB         244.529      4.313  56.697   <2e-16 ***
VarietyC         240.021      4.313  55.651   <2e-16 ***
FieldII          251.774      4.313  58.377   <2e-16 ***
VarietyB:FieldII   1.308      6.099   0.215    0.830
VarietyC:FieldII   3.345      6.099   0.548    0.584
```

Two-ways ANOVA - comparing three varieties in two fields - model control

```
> # Verifying the normality assumption
> Residuals <- residuals(Minter, "response")
> library(car)
> qqPlot(Residuals)
[1]  34 164
> shapiro.test(Residuals)

        Shapiro-Wilk normality test

data:  Residuals

W = 0.99704, p-value = 0.8602
```

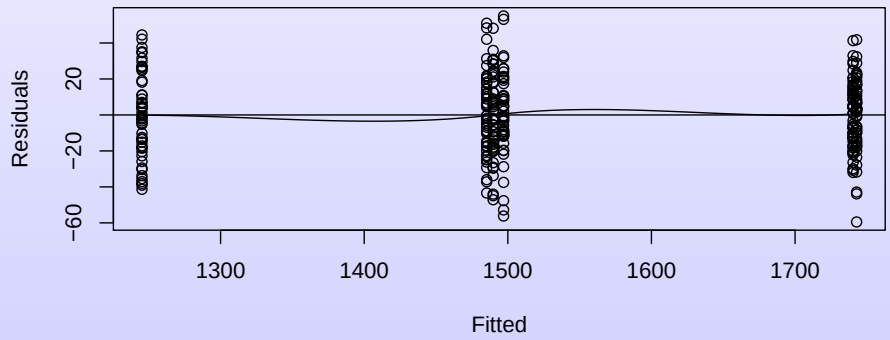
Two-ways ANOVA - comparing three varieties in two fields - model control

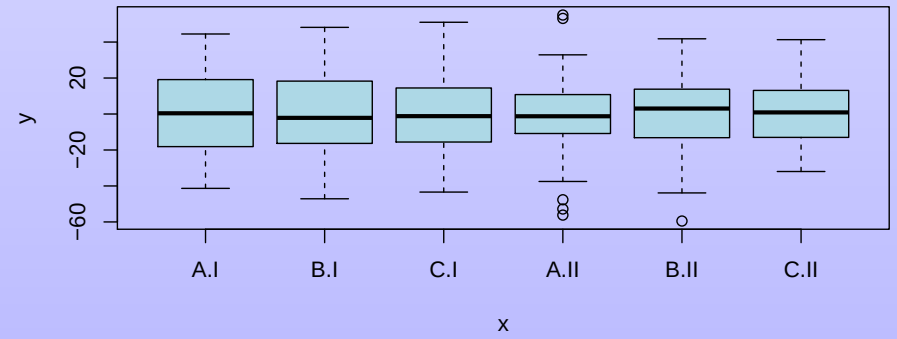


Two-ways ANOVA - comparing three varieties in two fields - model control

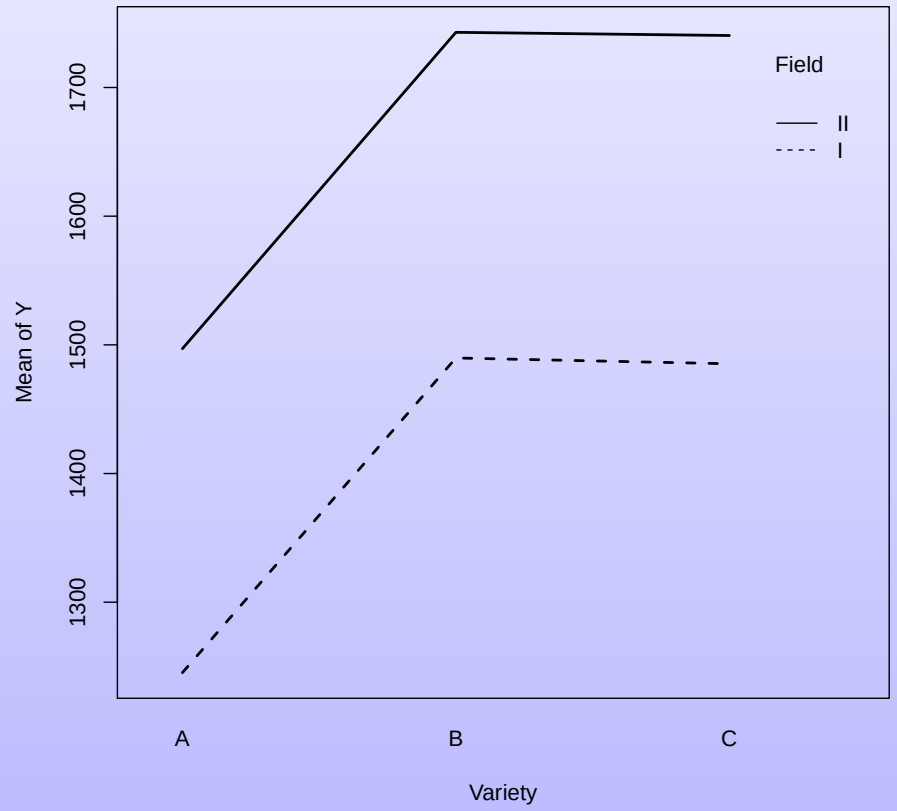
```
> # Verifying the variance homogeneity assumption
> Fitted <- fitted(Minter)
> par(mfrow = c(2,1))
> scatter.smooth(Fitted, Residuals); abline(h=0)
> plot(interaction(D$Variety, D$Field ), Residuals, col = "lightblue")
> par(mfrow = c(1,1))
> bartlett.test(Residuals, g = interaction(D$Variety, D$Field ))

        Bartlett test of homogeneity of variances

data:  Residuals and interaction(D$Variety, D$Field)

Bartlett's K-squared = 4.8859, df = 5, p-value = 0.43
```

Two-ways ANOVA - comparing three varieties in two fields - model control





Two-ways ANOVA - comparing three varieties in two fields - investigating additivity



Two-ways ANOVA - comparing three varieties in two fields - the additive model

- Y_{vfb} is the random variable representing the averaged weight of the b^{th} batch ($b = 1, ..., 50$) of the v^{th} variety ($v = A, B, C$) from the f^{th} field ($f = I, II$)

- The model assumes that the random variables $Y_{AI1}, \ldots, Y_{CII50}$ are:
  - independent,
  - normally distributed
  - have the same variance (say $Var(Y_{vfb}) = \sigma^2$)
  - The expectation can be written as a sum of a quantity depending on the variety and a quantity depending on the field

    (say $E(Y_{vfb}) = \tau_v + \beta_f$)

- In short,

$$Y_{vfb} \sim N\left(\tau_v + \beta_f, \sigma^2\right), \text{ for } v = A, B, C, \ f = I, II \text{ and } b = 1, ..., 50,$$

where $Y_{AI1}, \ldots, Y_{CII50}$ are independent.

Two-ways ANOVA - comparing three varieties in two fields - additive model

```
> # Fitting and testing an additive model
> Madd <- glm(Y ~ Variety + Field + 0, data = D)
> anova(Madd, Minter, test = "F")
Analysis of Deviance Table
Model 1: Y ~ Variety + Field + 0
Model 2: Y ~ Variety + Field + Variety:Field
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1       296     136863
2       294     136721  2   142.09 0.1528 0.8584
```

Two-ways ANOVA - comparing three varieties in two fields - additive model

```
> summary(Madd)

Call: glm(formula = Y ~ Variety + Field + 0, data = D)

...

Coefficients:

         Estimate Std. Error t value Pr(>|t|)
VarietyA 1244.482      2.483   501.2   <2e-16 ***
VarietyB 1489.665      2.483   600.0   <2e-16 ***
VarietyC 1486.175      2.483   598.6   <2e-16 ***
FieldII   253.326      2.483   102.0   <2e-16 ***
```

Two-ways ANOVA - comparing three varieties in two fields - testing the effect of variety

```
> MnoVar   <- glm(Y ~ Field, family = gaussian(link = "identity"), data = D)
```

```
> anova(MnoVar, Madd, test = "F")
```

```
Analysis of Deviance Table
```

```
Model 1: Y ~ Field
```

```
Model 2: Y ~ Variety + Field + 0
```

```
  Resid. Df Resid. Dev Df Deviance    F    Pr(>F)
```

```
1       298    4088285
```

```
2       296     136863  2  3951422 4273 < 2.2e-16 ***
```

Two-ways ANOVA - comparing three varieties in two fields - testing the effect of field

```
> MnoField <- glm(Y ~ Variety, family = gaussian(link = "identity"), data = D)
```

```
> anova(MnoField, Madd, test = "F")
```

```
Analysis of Deviance Table
```

```
Model 1: Y ~ Variety
```

```
Model 2: Y ~ Variety + Field + 0
```

```
  Resid. Df Resid. Dev Df Deviance     F    Pr(>F)
```

```
1       297    4949906
```

```
2       296     136863  1  4813043 10409 < 2.2e-16 ***
```

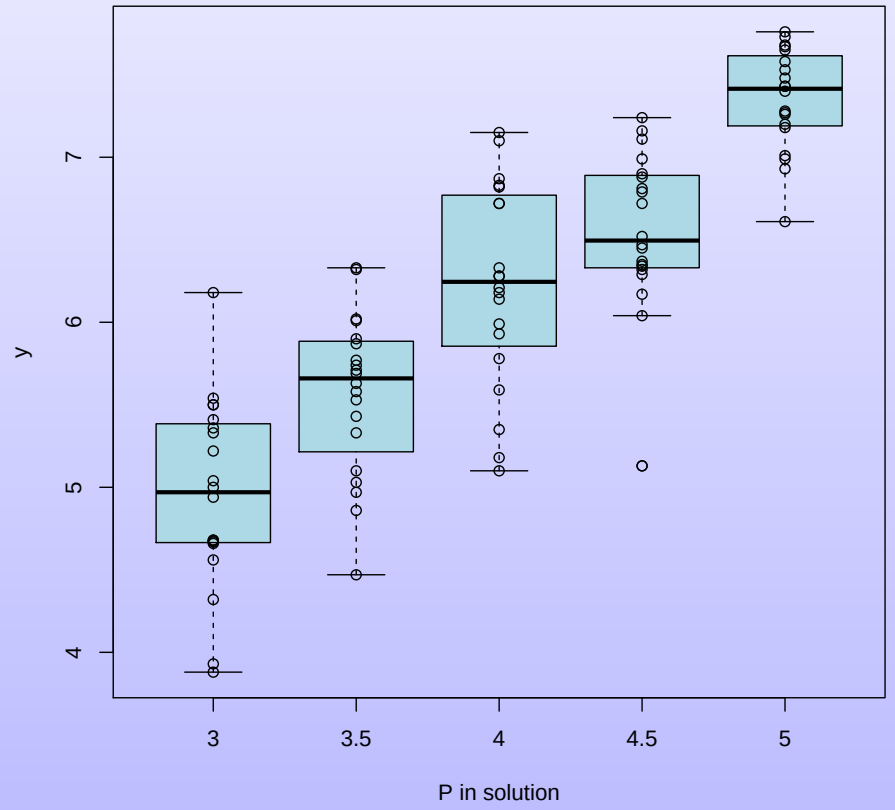
Two-ways ANOVA - comparing three varieties in two fields - Concluding

- We illustrated the classic models of one- and two-ways Gaussian classification models
  (one- and two-ways variance analysis models)
- The use of the normal distribution was justified by the central limit theorem
  (visible in this example)
- After postulating Gaussian models, we made some basic model check
- We concluded for an additive model with effect of both variety and field

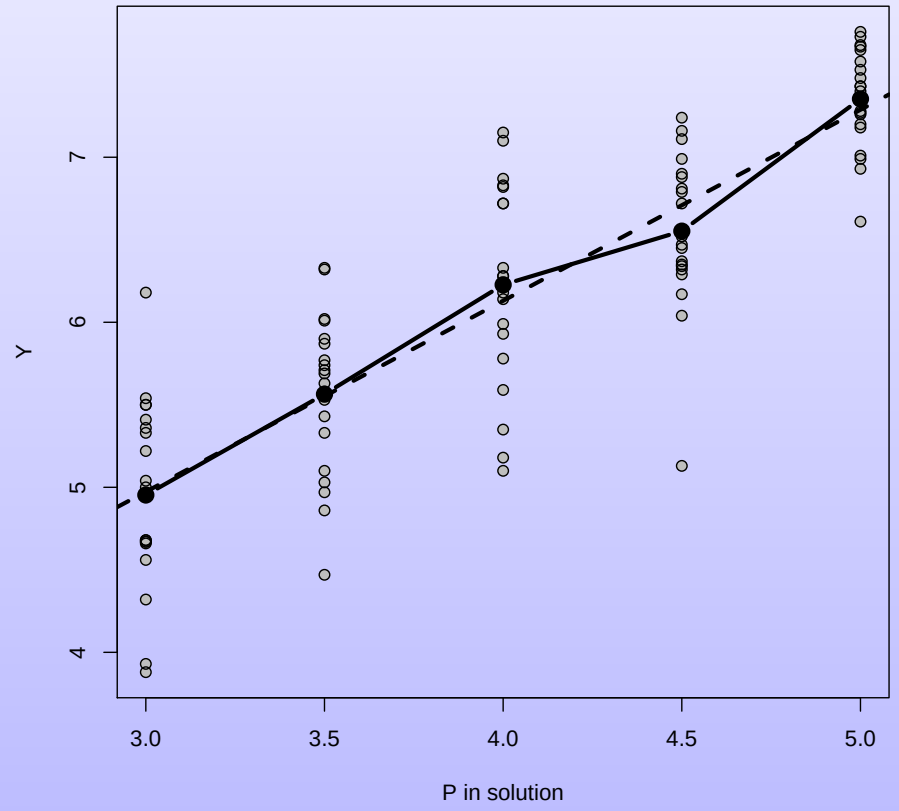
Linear Regression

- Maize cultivated in hydroponic solution
- 3, 3.5, 4, 4.5, 5 ppm P in solution
- 20 repetitions
- Registered the leaves weight after 10 days

Linear Regression



Linear Regression



- Y_{pr} weight of the r-th repetition
  subject to the amount p of Phosphorous
- We assume that the expected weight depends linearly on the amount of Phosphorous
- In symbols

$$E(Y_{pr}) = \alpha + \beta p$$

- We assume, moreover, that Y_{pr} is normally distributed with constant variance and
  that the observations are independent

Maize data - Linear regression

```
> linear <- glm(Y ~ Psol, family = gaussian, data = D)
> summary(linear)
Call: glm(formula = Y ~ Psol, family = gaussian, data = D)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.49990    0.28901    5.19 1.14e-06 ***
Psol         1.15760    0.07115   16.27  < 2e-16 ***
---
```

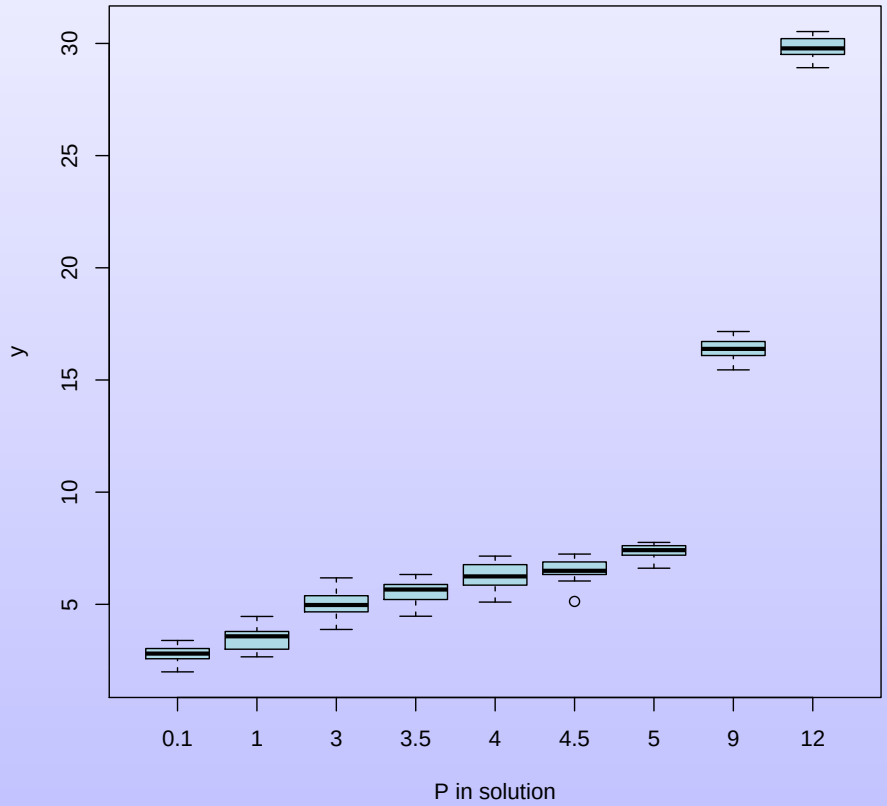
Maize data - Settting a Free curve model

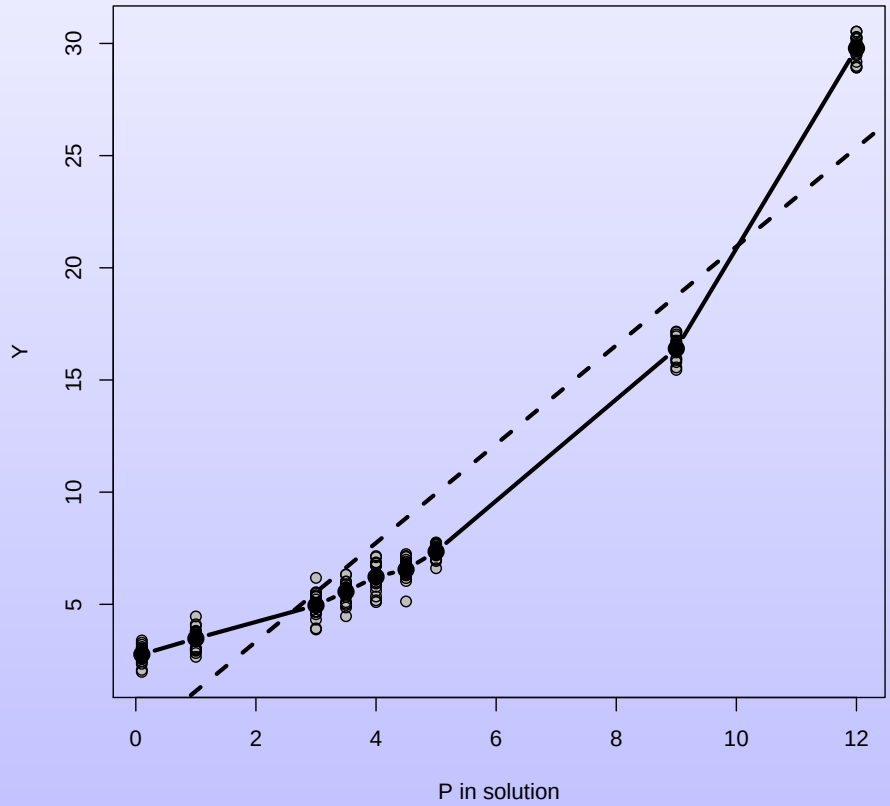
```
> free <- glm(Y ~ 0 + factor(Psol), family=gaussian, data = D)
> summary(free)
Call: glm(formula = Y ~ 0 + factor(Psol), family = gaussian, data = D)
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
factor(Psol)3     4.9535     0.1125   44.05   <2e-16 ***
factor(Psol)3.5   5.5645     0.1125   49.48   <2e-16 ***
factor(Psol)4     6.2275     0.1125   55.38   <2e-16 ***
factor(Psol)4.5   6.5525     0.1125   58.27   <2e-16 ***
factor(Psol)5     7.3535     0.1125   65.39   <2e-16 ***
```

Maize data - Testing linearity

```
> anova(linear, free, test = "F")
Analysis of Deviance Table

Model 1: Y ~ Psol
Model 2: Y ~ 0 + factor(Psol)
  Resid. Df Resid. Dev Df Deviance     F Pr(>F)
1        98     24.805
2        95     24.029  3  0.77625 1.023 0.3861
```

- In fact it was used more levels of P in the solution
- **0.1, 1 ,** 3, 3.5, 4, 4.5, 5, **9, 12** ppm





Maize data - Non-Linear Regression

```
> linear <- glm(Y ~ Psol, family = gaussian, data = D)
> free <- glm(Y ~ 0 + factor(Psol), family=gaussian, data = D)
> anova(linear, free, test = "F")
Analysis of Deviance Table
Model 1: Y ~ Psol
Model 2: Y ~ 0 + factor(Psol)
  Resid. Df Resid. Dev Df Deviance      F    Pr(>F)
1       178    1229.26
2       171      40.49  7   1188.8 717.22 < 2.2e-16 ***
```

Maize data - Setting a Free-curve Model

```
> summary(free)

Call: glm(formula = Y ~ 0 + factor(Psol), family = gaussian, data = D)

Coefficients:

                Estimate Std. Error t value Pr(>|t|)
factor(Psol)0.1   2.7690     0.1088   25.45   <2e-16 ***
factor(Psol)1     3.4790     0.1088   31.97   <2e-16 ***
factor(Psol)3     4.9535     0.1088   45.52   <2e-16 ***
factor(Psol)3.5   5.5645     0.1088   51.14   <2e-16 ***
factor(Psol)4     6.2275     0.1088   57.23   <2e-16 ***
factor(Psol)4.5   6.5525     0.1088   60.22   <2e-16 ***
factor(Psol)5     7.3535     0.1088   67.58   <2e-16 ***
factor(Psol)9    16.3985     0.1088  150.71   <2e-16 ***
factor(Psol)12   29.7840     0.1088  273.73   <2e-16 ***
```

An Exponential Model

- Y_{pr} weight of the r-th repetition
  subject to the amount p of Phosphorous
- We assume that

$$\log\left(E\left(Y_{pr}\right)\right) = \alpha + \beta p$$

  or equivalently,

$$E\left(Y_{pr}\right) = \exp\left(\alpha + \beta p\right)$$

- We assume, moreover, that Y_{pr} is normally distributed with constant variance and
  that the observations are independent

Maize data - An Exponential Model

```
> exponential <- glm(Y ~ Psol, family=gaussian(link = "log"), data = D)
> summary(exponential)
Call: glm(formula = Y ~ Psol, family = gaussian(link = "log"), data = D)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.0108650  0.0104704   96.55   <2e-16 ***
Psol        0.1985585  0.0009899  200.58   <2e-16 ***
```

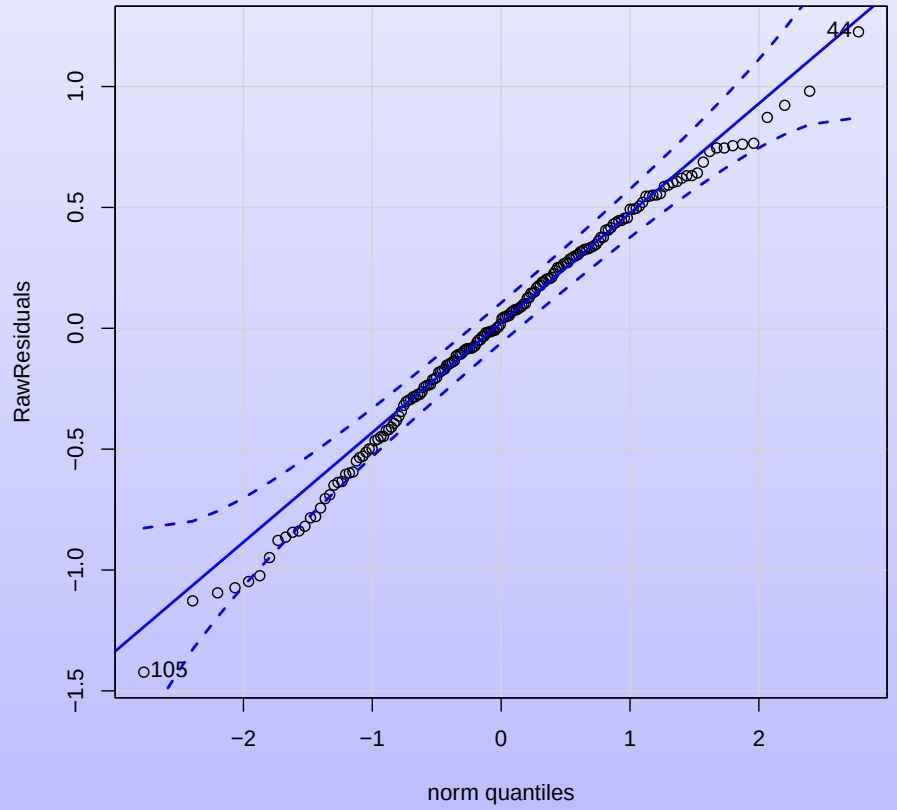
Maize data - Testing Adequacy of the Exponential Model

```
> exponential <- glm(Y ~ Psol, family=gaussian(link = "log"), data = D)
> anova(exponential, free, test = "F")
Analysis of Deviance Table
Model 1: Y ~ Psol
Model 2: Y ~ 0 + factor(Psol)
  Resid. Df Resid. Dev Df Deviance      F Pr(>F)
1       178     41.972
2       171     40.490  7   1.4827 0.8945 0.5121
```

Maize data - Testing Normality

```
> free <- glm(Y ~ 0 + factor(Psol), family=gaussian, data = D)
> RawResiduals <- residuals(free, "response")
> qqPlot(RawResiduals)
> shapiro.test(RawResiduals)
        Shapiro-Wilk normality test
data:  RawResiduals
W = 0.99024, p-value = 0.2573
```

Checking the normality assumption



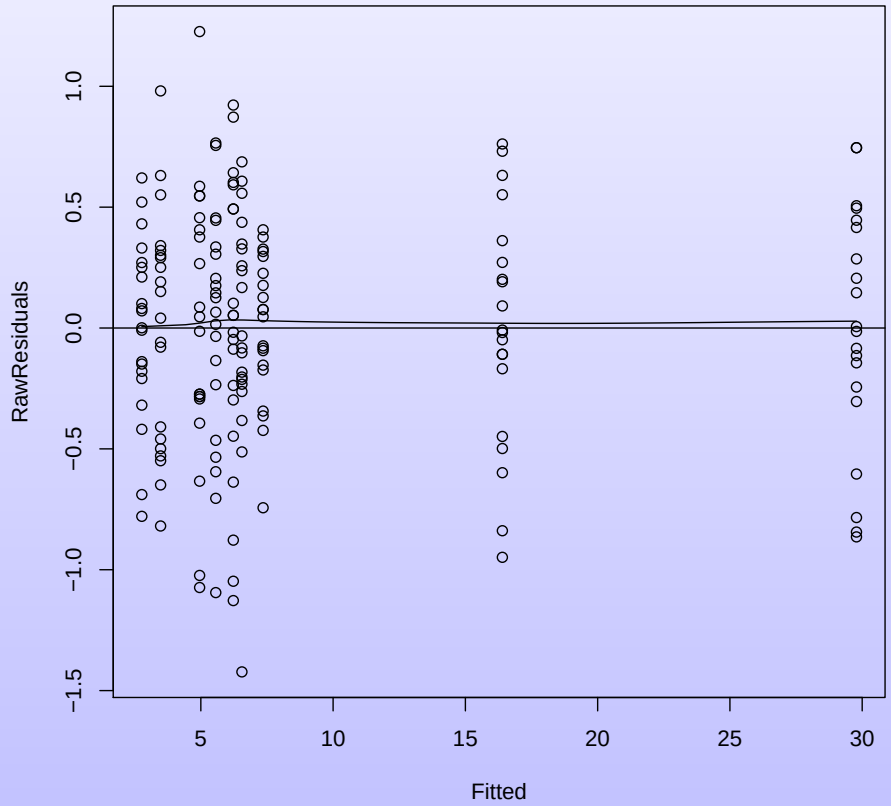
Maize data - Homocedasticity (variance homogeneity)

```
> Fitted <- fitted(free)
```

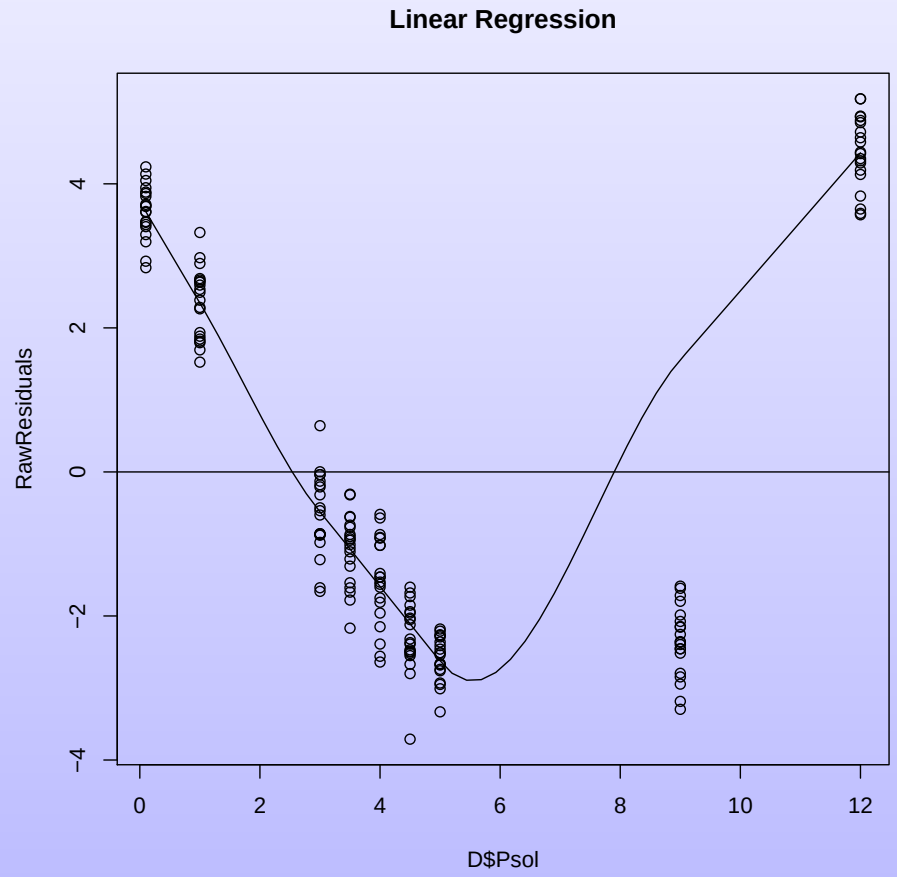
```
> scatter.smooth(Fitted, RawResiduals)
```

```
> scatter.smooth(Fitted, RawResiduals); abline(h=0)
```

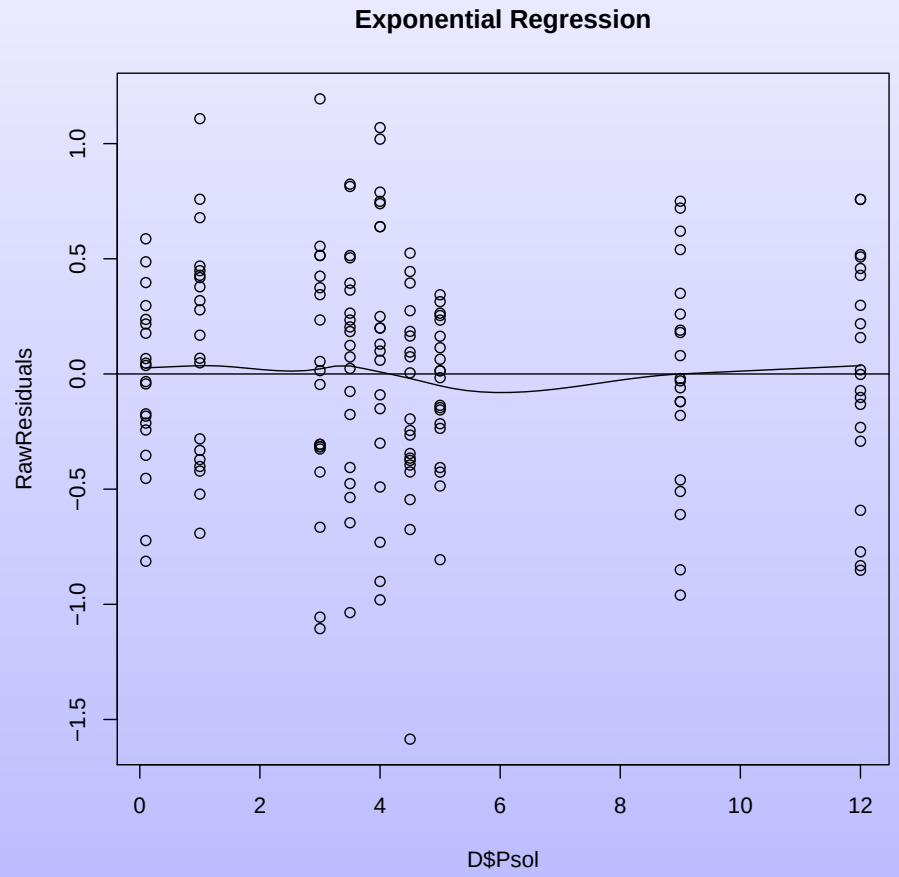
```
> bartlett.test(Y ~ factor(Psol),data=Ch5.maize.ALL)

        Bartlett test of homogeneity of variances

data:  Y by factor(Psol)

Bartlett's K-squared = 11.767, df = 8, p-value = 0.1619
```



Maize data - Verifying the Adequacy of the Linear Model



Verifying the Adequacy of the Exponential Model



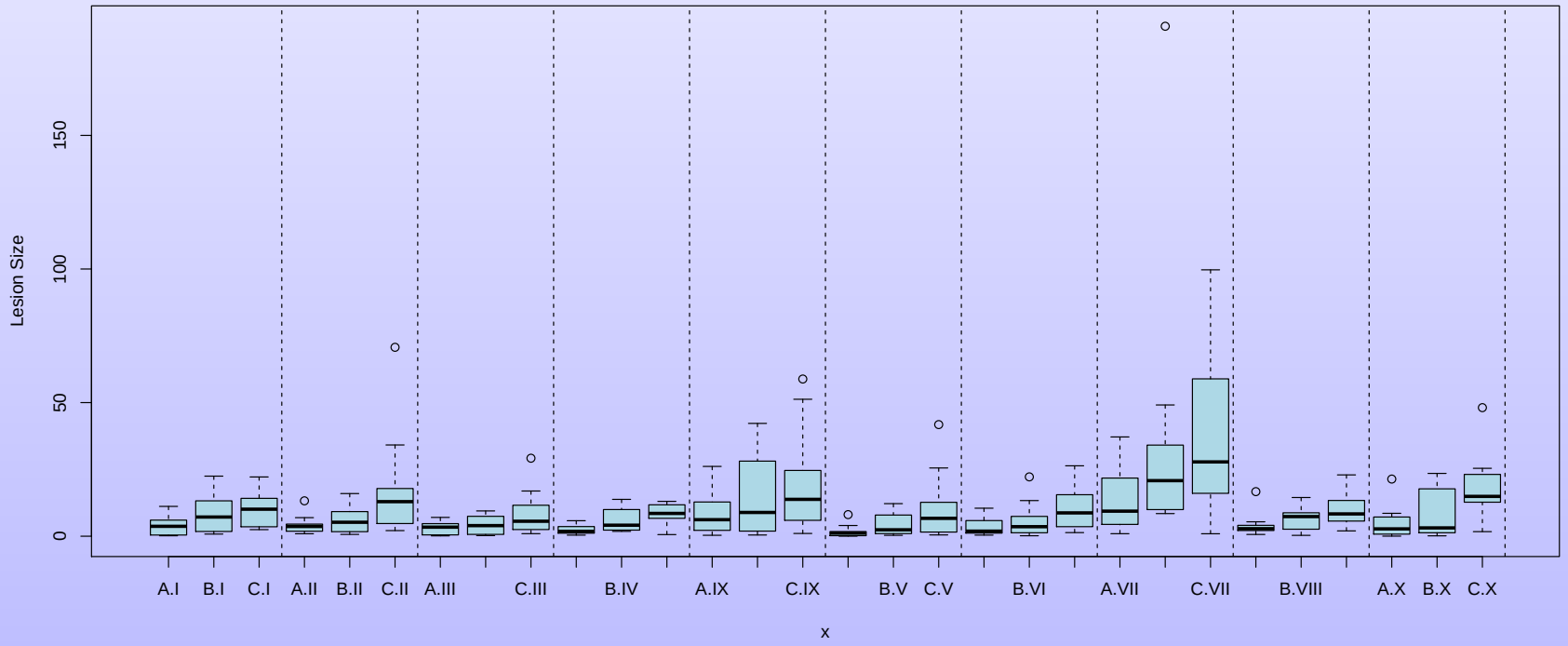
Linear and Non-linear Gaussian Regression - Concluding

- We demonstrated how to construct and use linear and non-linear Gaussian regression models
- It is possible to use the function "lm" instead of "glm", but then there is no possibility to specify the link function

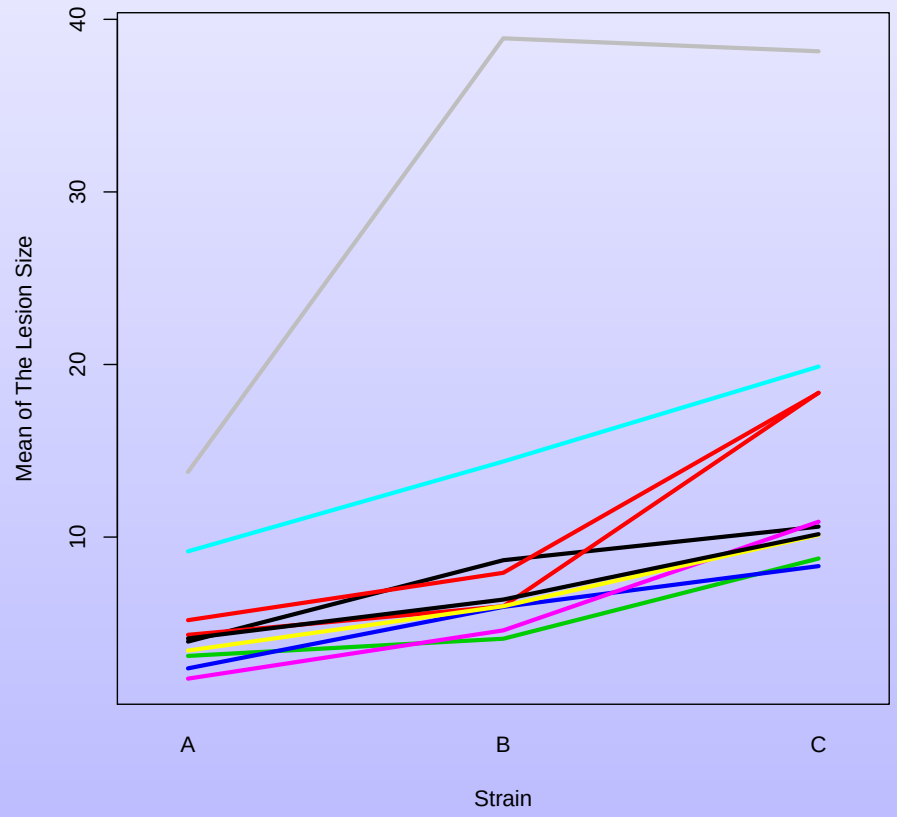
Initial Example of Non-Gaussian Models - Fungal resistance essay

- Several measurements of fungal resistance in a cultivated plant
- Three fungal strains: A, B and C.
- 10 plants, 10 repetitions (leaves) inoculated
- Responses:
  Lesion size
- Different leaves used for the three determinations
- We analyse the lesion sizes in detail

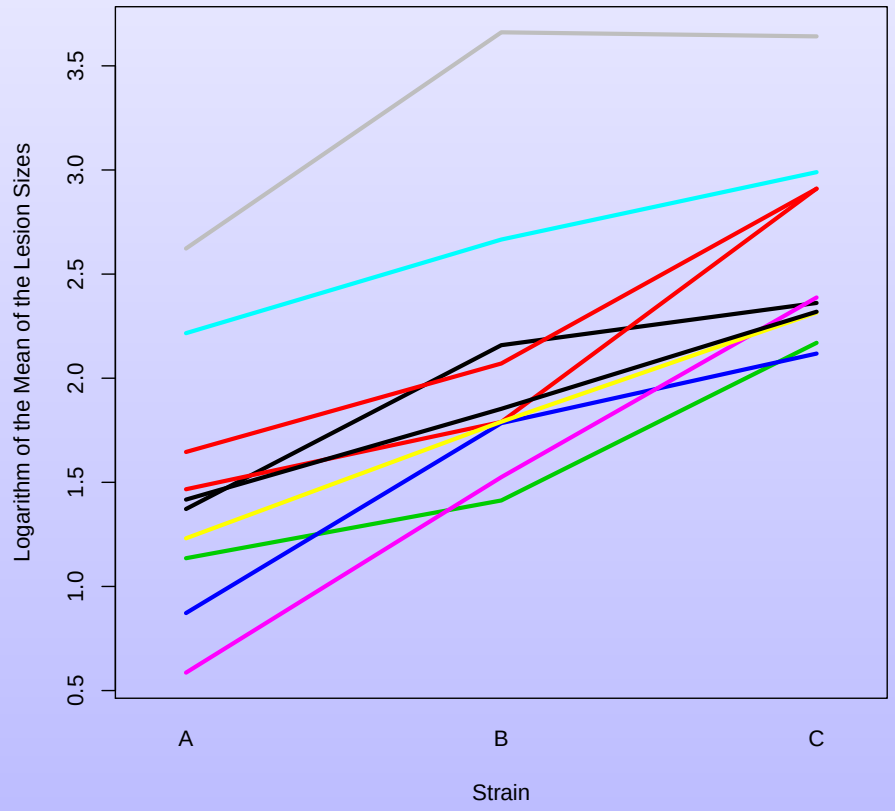
Initial Example of Non-Gaussian Models - Fungal resistance essay



Initial Example of Non-Gaussian Models - Fungal resistance essay



Initial Example of Non-Gaussian Models - Fungal resistance essay



A Gaussian Linear Model - the naive approach ...

- Denote by $\mathcal{Y}_{b,t,r}$ the random variable representing the lesion size of the r^{th} replicate ($r = 1, \ldots, 10$) of the experimental units of the b^{th} plant (or cluster, $b = I, \ldots, X$) that received the t^{th} strain ($t = A, B, C$).
- $\mathcal{Y}_{I,A,1}, \ldots, \mathcal{Y}_{X,C,10}$ are independent and normally distributed and for $b = I, \ldots, X$, $t = A, B, C$ and $r = 1, \ldots, 10$,

$$\log\left\{E\left(\mathcal{Y}_{btr}\right)\right\} = \tau_t + \beta_b\,,$$

or equivalently,

$$E\left(\mathcal{Y}_{btr}\right) = \exp\left(\tau_t + \beta_b\right) = \exp(\tau_t)\exp(\beta_b)\,.$$

- First, consider a model with effect modification (or interaction) where

$$E\left(\mathcal{Y}_{btr}\right) = \exp\left(\gamma_{tb}\right)\,.$$

A Gaussian Linear Model - the naive approach ...

```
> library(GLMMstudy)
```

```
> data("FungusResistance"); D <- FungusResistance
```

```
> str(D)
```

```
'data.frame':        300 obs. of  5 variables:
 $ Counts    : num  1 1 0 2 3 2 2 2 2 0 ...
 $ Plant     : Factor w/ 10 levels "I","II","III",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Strain    : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
 $ HyperSens : num  11 7 14 3 12 9 11 7 8 5 ...
 $ LesionSize: num  0.44 6.02 3.74 7.92 4.58 ...
```

A Gaussian Linear Model - the naive approach ...

```
> M <- glm(LesionSize ~ Strain * Plant, family = gaussian(link = "log") ,data = D)
```

```
> Raw_Residuals <- residuals(M, "response")
```

```
> library(car)
```

```
> qqPlot(Raw_Residuals)
```

```
[1] 192 203
```

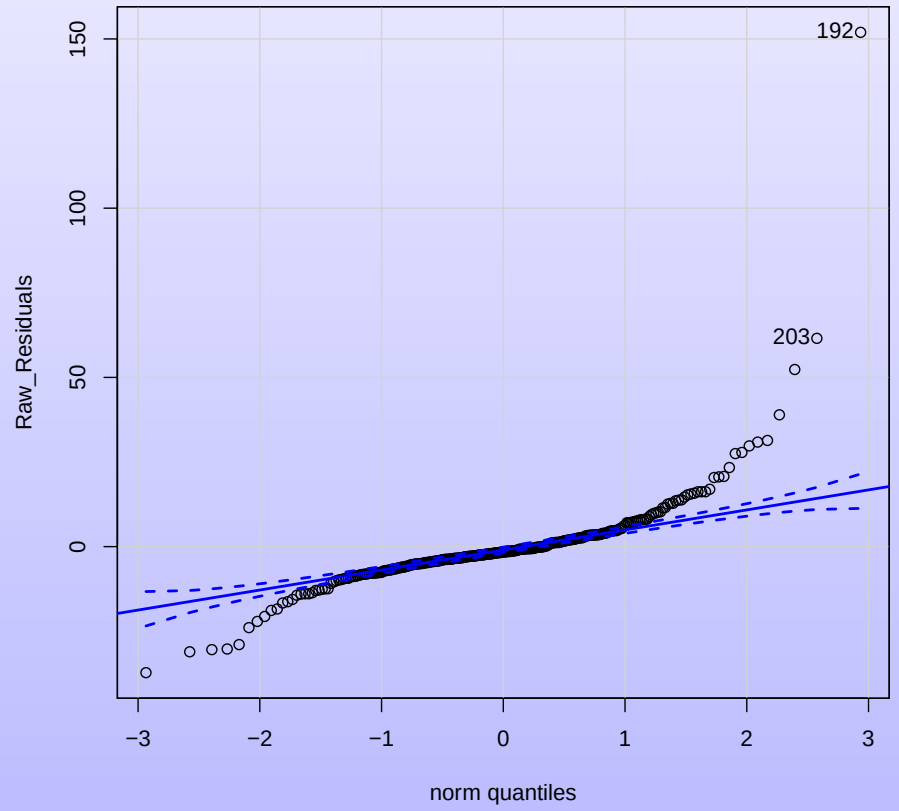
```
> shapiro.test(Raw_Residuals)
```

```
        Shapiro-Wilk normality test
```

```
data:  Raw_Residuals
```

```
W = 0.65676, p-value < 2.2e-16
```

A Gaussian Linear Model - the naive approach ...



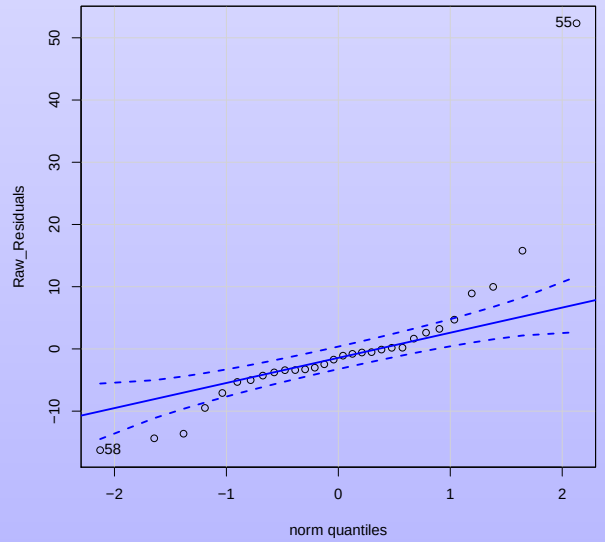
A Gaussian Linear Model - the naive approach ...

```
> plot(interaction(D$Strain, D$Plant), Raw_Residuals, col = "lightblue")
```

```
> bartlett.test(Raw_Residuals, g = interaction(D$Strain, D$Plant))
```

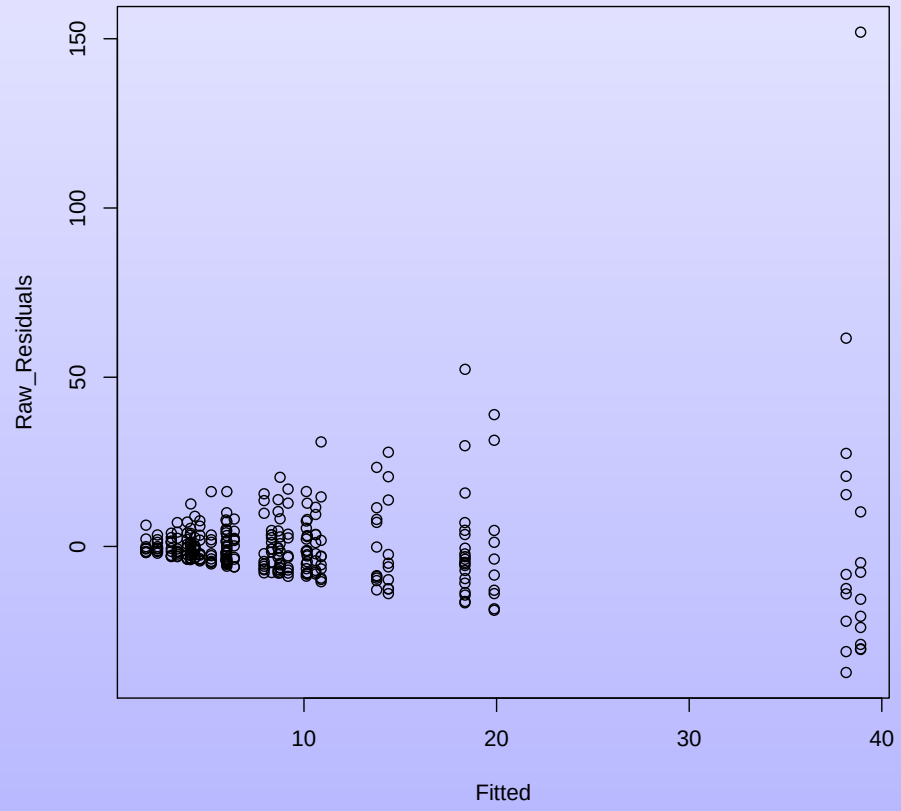
```
        Bartlett test of homogeneity of variances
```

```
Bartlett's K-squared = 374.67, df = 29, p-value < 2.2e-16
```



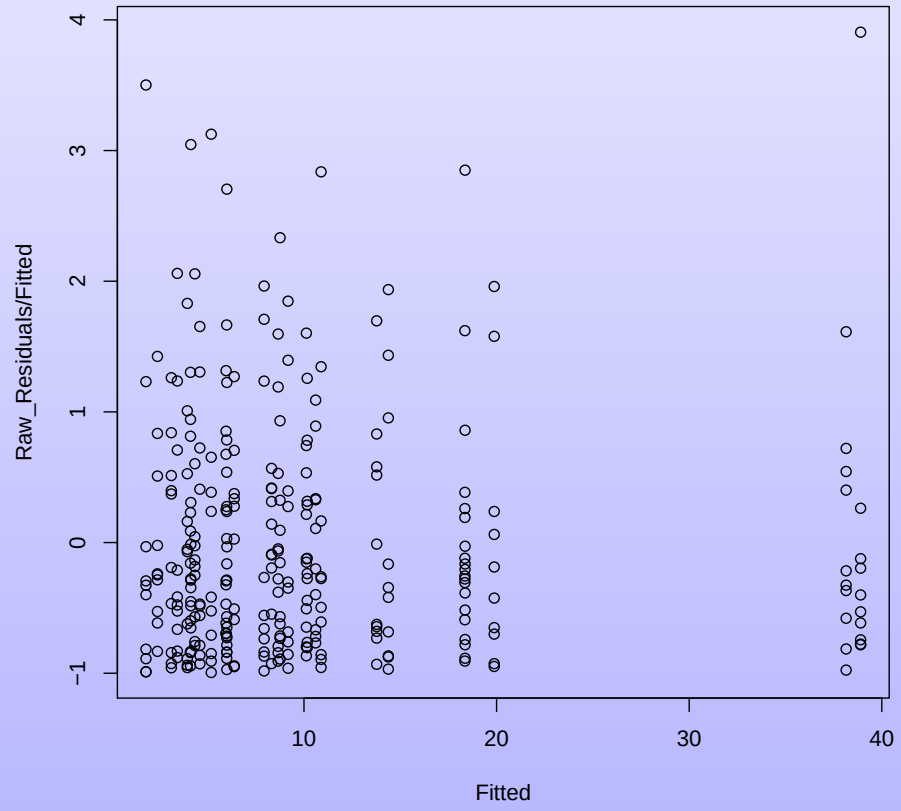
A Gaussian Linear Model - the naive approach ...

```
Fitted <- fitted(M); plot(Fitted, Raw_Residuals)
```



A Gaussian Linear Model - the naive approach ...

```
plot(Fitted, Raw_Residuals/Fitted)
```



A Gaussian Linear Model - the naive approach ...

- The gaussian linear model is not adequate for two reasons
- First, the responses are not normally distributed
- Second, the observations are probably not independent
  Several observations taken from the same plant ...
- Solution: There are indications that a Gamma distribution might be suitable
- Solution: make 10 separate analyses, one for each plant
  What a limitation!

Initial Example of Non-Gaussian Models

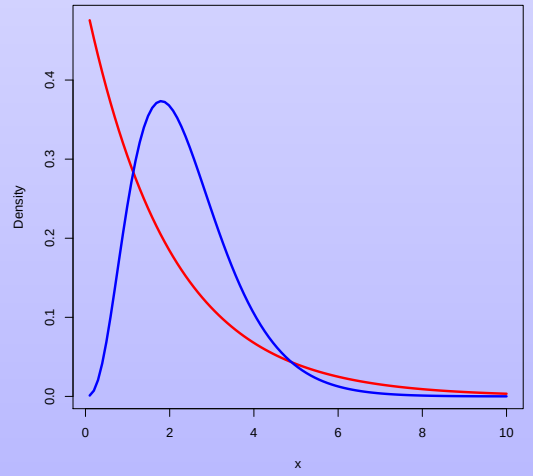
- We will use a GLM (and a GLMM on some weeks) defined with the gamma distribution
- The model will contain a factor representing the effect of the strains and we will make separate analyses per plant.
- On some weeks, we will work with a model will containing a fixed effect representing the effect of the strains and a random component representing the plant.
- But before we present some basic results on the Gamma distribution.

The Gamma Distribution - Definition

- A probability distribution on the positive real numbers with probability density of the form, for $\alpha > 0$ and $\beta > 0$,

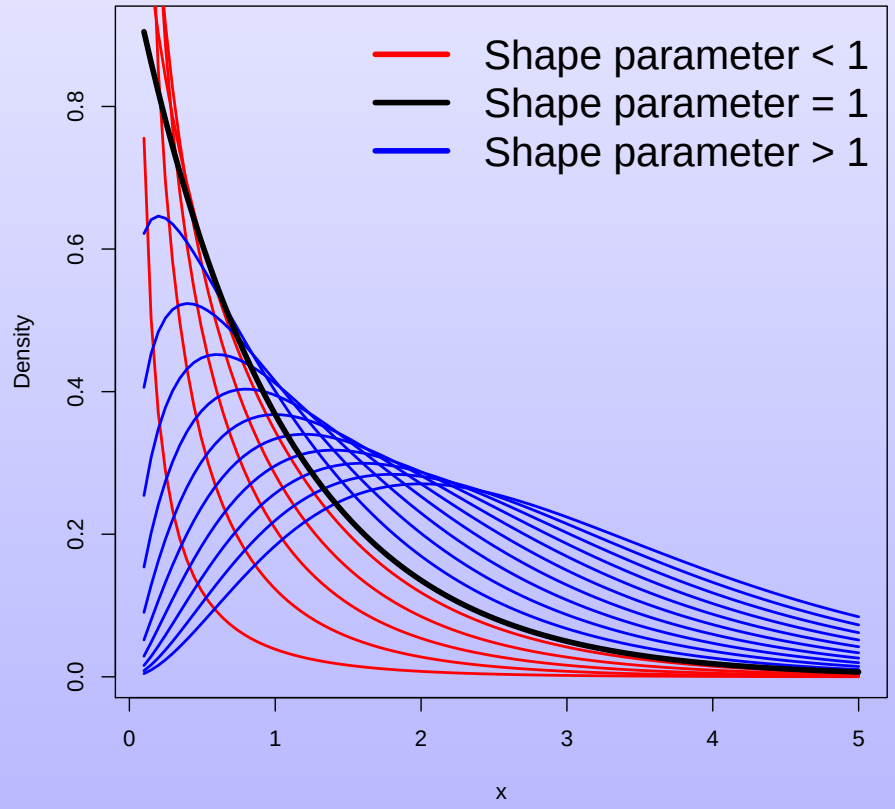
$$p(y; \alpha, \beta) = y^{\alpha-1} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \exp\left(-y/\beta\right), \text{ for } y > 0,$$

  is said to be a *Gamma distribution*. Notation $X \sim G(\alpha, \beta)$
- The parameters $\alpha > 0$ and $\beta > 0$ are called the *shape* and the *scale* parameters, respectively.



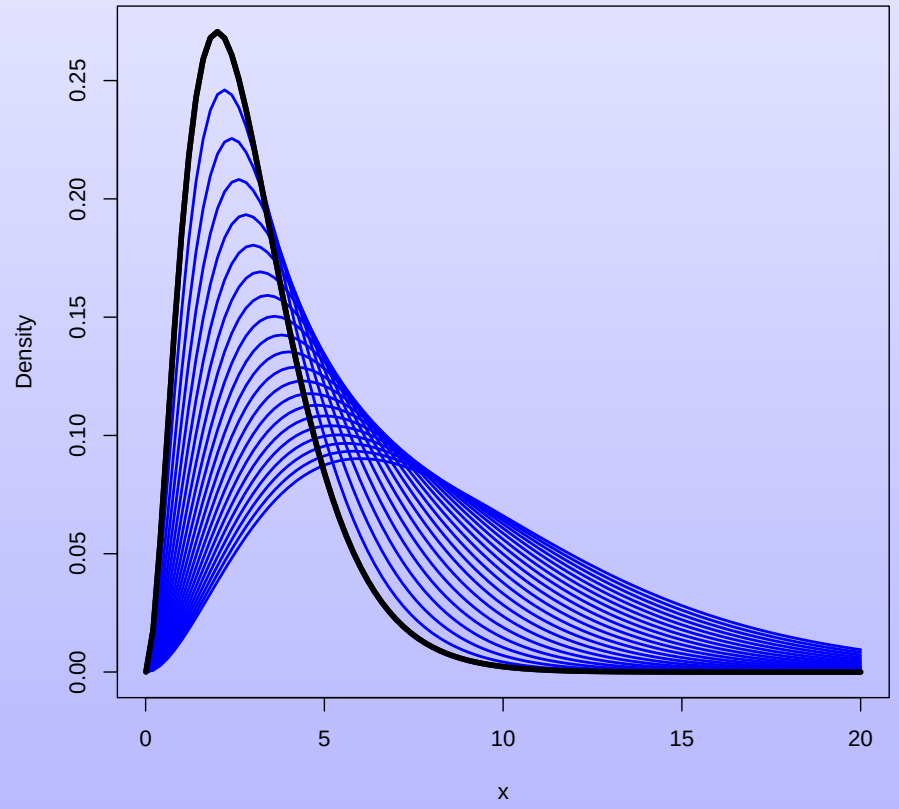
The Gamma Distribution

Changing the shape parameter changes the form of the density



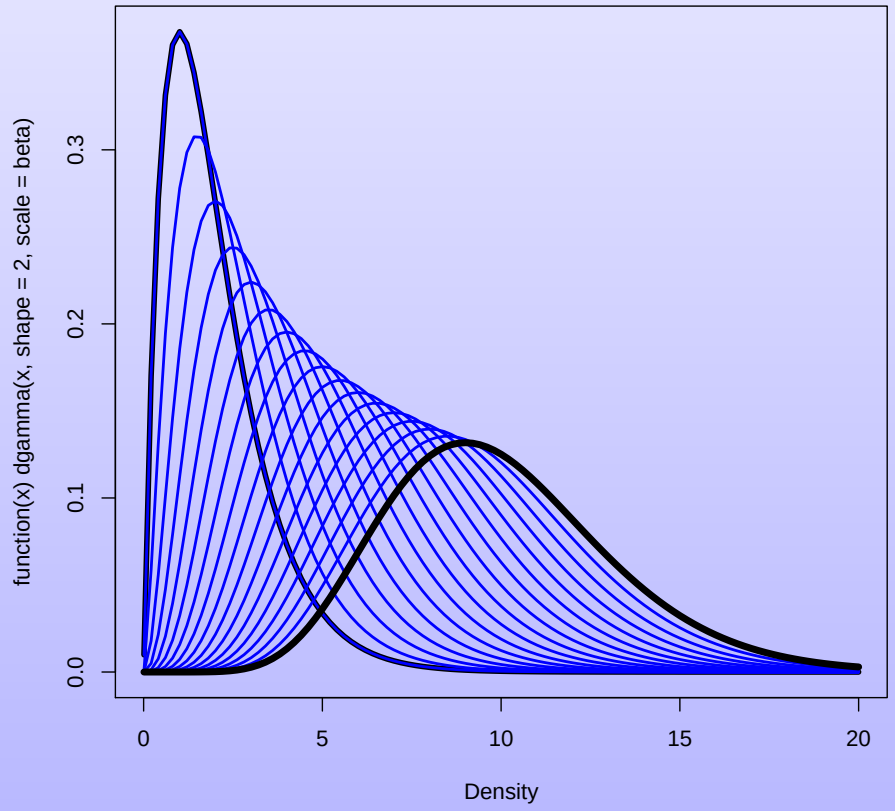
The Gamma Distribution

Changing the scale parameter re-scale the density



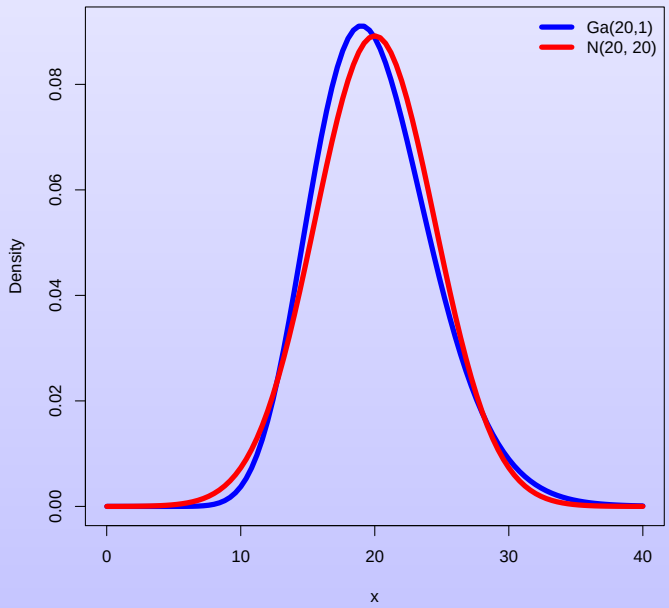
The Gamma Distribution

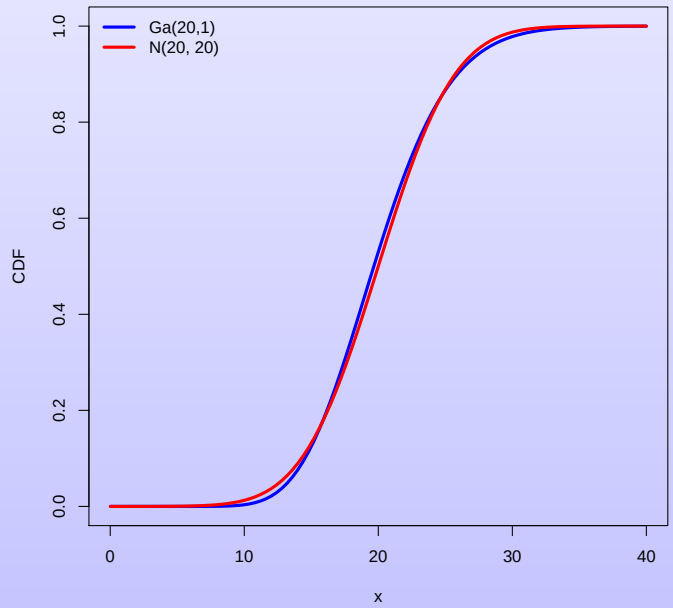
Increasing the shape parameters decreases the right-skewness



The Gamma Distribution

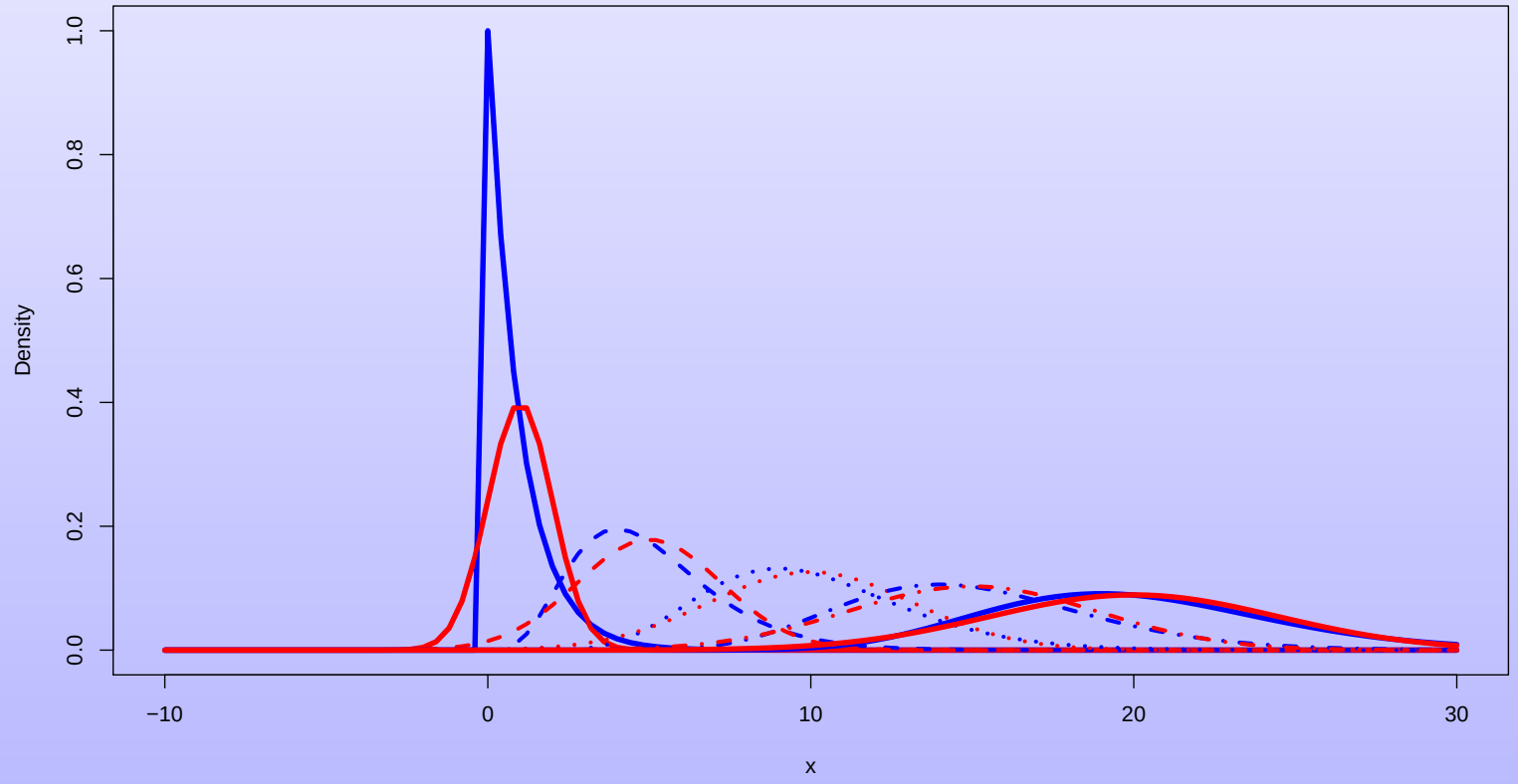
The Gamma distribution can mimic the normal distribution!





The Gamma Distribution

The Gamma distribution converges to the normal distribution as $\alpha \longrightarrow \infty$



The Gamma Distribution - Basic facts

- $p(y;\alpha,\beta) = y^{\alpha-1}\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\exp\left(-y/\beta\right)$, for $y > 0$,
- Notation $X \sim G(\alpha,\beta)$
- If $X \sim G(\alpha,\beta)$ then $E(X) = \alpha\beta$ and $Var(X) = \alpha\beta^2$.
- The skewness of X is $2/\sqrt{\alpha}$,

  implying that the skewness of Gamma distributions can be made arbitrarily small by choosing values of the shape parameter large enough.

- The moment generating function and the characteristic function of the Gamma distribution with shape and scale parameters α and β, respectively, are $M(t) = (1-\alpha t)^{-\beta}$ (for $t > 1/\beta$) and $\varphi(t) = (1-\alpha it)^{-\beta}$ (for t real). Differentiating the moment-generating function or the characteristic function yields the moments of the Gamma distribution of all orders.

The Gamma Distribution - Basic facts

- The family of distributions formed by the Gamma distribution is a dispersion model

  A dispersion model generated by the unit deviance $d(y;\mu) = 2\left\{-\log(y/\mu) + \frac{y-\mu}{\mu}\right\}$, where $y > 0$ and $\mu > 0$.

- The Gamma distributions form an exponential dispersion model with unit variance function $V(\mu) = \mu^2$

  An exponential dispersion model with canonical parameter $\theta = -1/\mu$ (where $\mu = \alpha\beta$) and moment generator $K(\theta) = -\log(\theta)$.

- Therefore, we can construct generalised linear models and generalised linear mixed models defined with Gamma distributions to model Gamma distributed responses.

- Due to the flexibility of the family of Gamma distributions, these models are expected to have a wide range of applicability.

The Gamma Distribution - Basic facts

- The Gamma distributions appear naturally in many applications for several reasons; three of them are given below.
- Sums of independent squares of normal distributed random variables are Gamma distributed

  (since the chi-square distributions are particular cases of Gamma distributions)
- The Erlang distributions (*i.e.* ., the sum of independent exponentially distributed random variables), which are the distributions of the waiting time until the arrivals in a Poisson process, are issues of the Gamma distribution.
- The gamma distribution is the maximum entropy probability distribution among the distributions taking positive values with a given expectation.

  Consequence: the Gamma distributions minimise the amount of prior information built into the distribution.

  Moreover, physical systems tend to move towards maximal entropy configurations.

A Gamma Generalised Linear Model - Defining a model

- Denote by $\mathcal{Y}_{b,t,r}$ the random variable representing the lesion size of the r^{th} replicate ($r = 1, \ldots, 10$) of the experimental units of the b^{th} plant (or cluster, $b = I, \ldots, X$) that received the t^{th} strain ($t = A, B, C$).
- $\mathcal{Y}_{I,A,1}, \ldots, \mathcal{Y}_{X,C,10}$ are independent and Gamma distributed and for $b = I, \ldots, X$, $t = A, B, C$ and $r = 1, \ldots, 10$,

$$\log\left\{E\left(\mathcal{Y}_{btr}\right)\right\} = \tau_t + \beta_b\,,$$

or equivalently,

$$E\left(\mathcal{Y}_{btr}\right) = \exp\left(\tau_t + \beta_b\right) = \exp(\tau_t)\exp(\beta_b)\,.$$

- But, observations arising from the same plant are **not** independent ...

... we will demonstrate that in the next lecture ...

A Gamma Generalised Linear Model - Defining a model

- We will work the data of only plant II
- Denote by $\mathcal{Y}_{t,r}$ the random variable representing the lesion size of the r^{th} replicate ($r = 1, \ldots, 10$) of the experimental units that received the t^{th} strain ($t = A, B, C$).
- $\mathcal{Y}_{A,1}, \ldots, \mathcal{Y}_{C,10}$ are independent and Gamma distributed and for $t = A, B, C$ and $r = 1, \ldots, 10$,

$$E\left(\mathcal{Y}_{tr}\right) = \tau_t \,,$$

A Generalised Linear Model

```
> D <- subset(FungusResistance, Plant == "II")
```

```
> str(D)
```

```
'data.frame':        30 obs. of  5 variables:
 $ Counts    : num  0 1 1 2 2 0 1 1 1 1 ...
 $ Plant     : Factor w/ 10 levels "I","II","III",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Strain    : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
 $ HyperSens : num  15 13 11 11 12 9 19 9 12 6 ...
 $ LesionSize: num  3.25 1.05 0.92 3.54 13.25 ...
```

A Naive Gaussian Generalised Linear Model

```
> M <- glm(LesionSize ~ Strain + 0, family = gaussian(link = "identity") ,data = D)
```

```
> Raw_Residuals <- residuals(M, "response")
```

```
> shapiro.test(Raw_Residuals)
```

```
        Shapiro-Wilk normality test
```

```
data:  Raw_Residuals
```

```
W = 0.70915, p-value = 2.126e-06
```

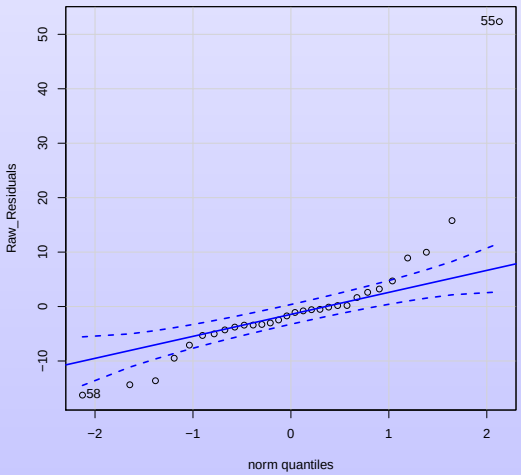
```
> bartlett.test(Raw_Residuals, g = D$Strain)
```

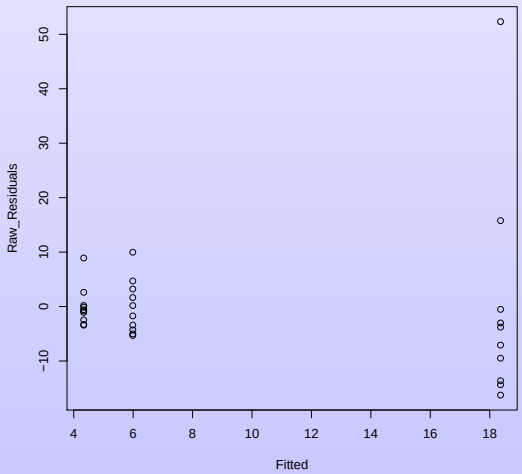
```
        Bartlett test of homogeneity of variances
```

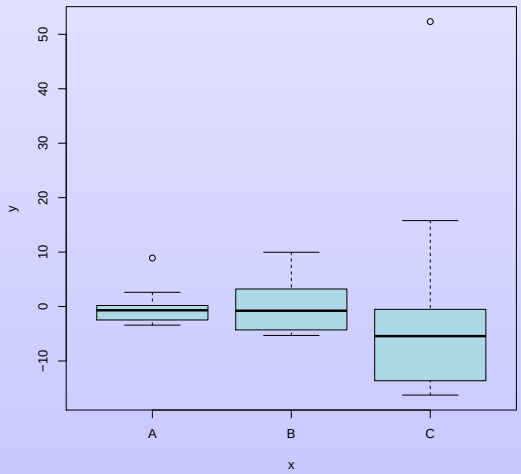
```
data:  Raw_Residuals and D$Strain
```

```
Bartlett's K-squared = 28.207, df = 2, p-value = 7.499e-07
```

A Naive Gaussian Generalised Linear Model







A Gamma Generalised Linear Model - fitting the model

```
> M <- glm(LesionSize ~ Strain + 0, family = Gamma(link = "identity") ,data = D)
```

```
> summary(M)
```

```
Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
```

```
-1.6056  -0.9660  -0.2088   0.2015   1.7331
```

```
Coefficients:
```

```
        Estimate Std. Error t value Pr(>|t|)
```

```
StrainA    4.335      1.284   3.377  0.00224 **
```

```
StrainB    5.988      1.773   3.377  0.00224 **
```

```
StrainC   18.364      5.439   3.377  0.00224 **
```

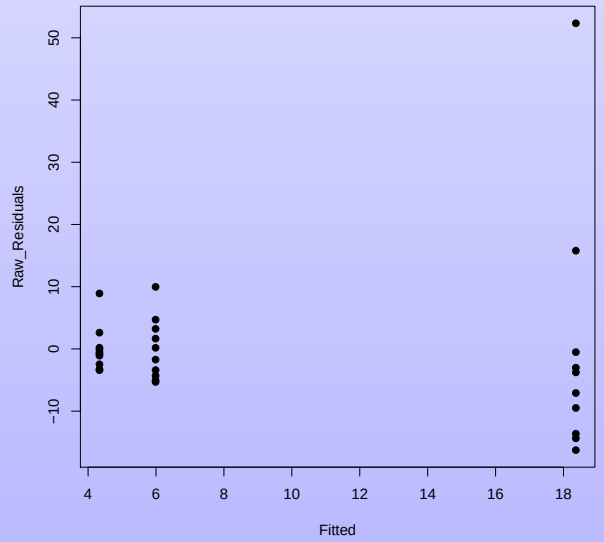
A Gamma Generalised Linear Model - Some control model

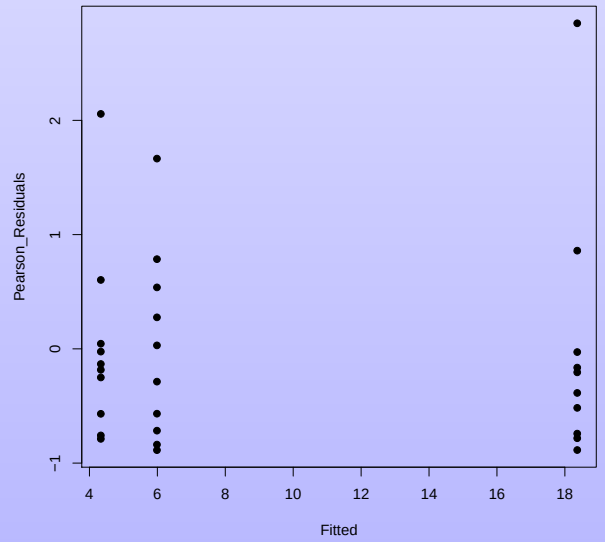
```
> Pearson_Residuals <- residuals(M,"pearson"); Fitted <- fitted(M)
```

```
> Raw_residuals <- residuals(M,"response")
```

```
> Fitted <- fitted(M)
```

```
> plot(Fitted, Raw_Residuals, pch = 19); plot(Fitted, Pearson_Residuals, pch = 19)
```





A Gamma Generalised Linear Model - testing

```
> M0 <- glm(LesionSize ~ 1, family = Gamma(link = "identity") ,data = D)
```

```
> anova(M0,M, test = "F")
```

```
Analysis of Deviance Table
```

```

```

```
Model 1: LesionSize ~ 1
```

```
Model 2: LesionSize ~ Strain + 0
```

```
  Resid. Df Resid. Dev Df Deviance     F   Pr(>F)
```

```
1        29     35.593
```

```
2        27     23.460  2   12.133 6.916 0.003757 **
```

```
---
```

Closing - Additional activities related to this lecture

- Study the program "Program-06-Lecture07-NormalModels" with the R-codes implementing the analyses performed here
- Run the tutorials: "Stat-Tutorial-04-TheNormalDistribution", "Stat-Tutorial-06-TheCentralLimitTheorem", "Stat-Tutorial-07-TheFailureOfTheCentralLimitTheorem", "Stat-Tutorial-11-AnscombeQuartet"
- Read the texts "Remarks on Model Definition" for a discussion on how to formulate a (one-way classification) model and "additional text on the corner point parametrisation"

  (both available in the work page in the section "lecture notes")