Basic Statistical Analysis

Rodrigo Labouriau

Department of Mathematics, Aarhus University

Module 3, Day 5 - Poisson Models - 2025
(regression structures)

1

General Remark

This material is only for internal use in the course.

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

Outline

2

Review:

Important general 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 regression and binomial covariance analysis models

Review:

Binomial models

- We made a distinction between response variable and explanatory variables
- Three categories of binomial models:
    - Pure discrete explanatory variables:
      one-way, two-ways, ..., k-ways
      (Exercises Ex-3-1 and Ex-3-2)
    - Pure continuous explanatory variables:
      logistic, probit, complementary-log-log regression
      (Exercise Ex-3-3)
    - Mixed type: variants of the covariance analysis type models
      (Exercise Ex-3-4)

Review:

Binomial models, one-way classification models

- One classification variable, say T
  Y_{ti} the i^{th} repetition of observations classified as t
- $Y_{11}, Y_{12}, \ldots$ independent
- $Y_{ti} \sim Bi(n_{ti}, p_t)$, for $t = 1, 2, \ldots$
- Equivalently
  $Y_{ti} \sim Bi(n_{ti}, p_{ti})$, for $t = 1, 2, \ldots$
  where $logit(p_{ti}) = T_t$

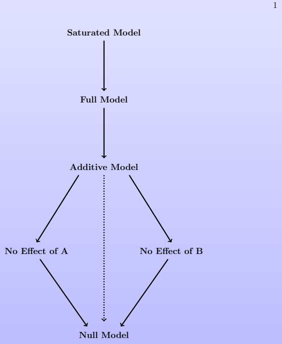
Review:

Binomial models, two-ways classification models

- Two classification variables, say T and S
  Y_{tsi} the i^{th} repetition of observations classified as t and s
- $Y_{111}, Y_{112}, \ldots$ independent
- $Y_{tsi} \sim Bi(n_{tsi}, p_{tsi})$, for $t, s = 1, 2, \ldots$
  with several possibilities for p_{tsi} (yielding different models)
- The possible models are:
    - $logit(p_{tsi})$ depends on t, s and i (the saturated model)
    - $logit(p_{tsi}) = (T * S)_{ts}$ (interaction model)
    - $logit(p_{tsi}) = T_t + S_s$ (additive model)
    - $logit(p_{tsi}) = S_s$ (no effect of T)
    - $logit(p_{tsi}) = T_s$ (no effect of S)
    - $logit(p_{tsi}) = K$, where K is a constant (null model)

Review:

Possible two-way classification models



Review:

Binomial models, regression

- $Y_{xi} \sim Bi(n_{xi}, p_{xi})$ where
  $g(p_{xi}) = \alpha + \beta x$,
  Here g is a function given (called the link function)
- Examples of functions link functions: logit, probit, identity, etc

Review:

Binomial models, multiple regression

- A classification variable, say T
  and one regression for each level of T
- $Y_{txi} \sim Bi(n_{xi}, p_{txi})$ where
  $g(p_{xi}) = \alpha_t + \beta_t x,$

  Here g is a function given (called the link function)
- Examples of functions link functions: logit, probit, identity, etc
- A range of structures arises

  (blackboard)

Review:

The general idea of the likelihood ratio test

- Idea:

  The reduction is reasonable when the "reduced model" fits the data as well as the "large model".
- Discrepancy of two models:

  Evaluate the likelihood function of both models at their maxima

  Examine the ratio of the two likelihood functions

  Large differences indicate discrepancy
- Values of this ratios close to 1 indicate that the two models are **not** "in disagreement"

Review:

The general idea of the likelihood ratio test

- Equivalently,
  evaluate the logarithm of the ratio of the two likelihood,
  examine the difference of the log-likelihood
- This difference is a positive quantity can be used to make tests
- The *log-likelihood ratio statistic* is defined by

$$\Lambda = 2\left\{l_L - l_R\right\} \quad ,$$

Review:

The general idea of the likelihood ratio test, example

- Large model containing the parameters $p_1, \ldots, p_t$

  Reduced model containing only the parameter p
- The log-likelihood ratio statistics is

$$\Lambda = 2\left\{ l_L(\hat{p}_1, \ldots, \hat{p}_t) - l_R(\hat{p}) \right\} = 2\left\{ l_L - l_R \right\} \quad ,$$

where $l_R(\hat{p}) = l_R$ and $l_L(\hat{p}_1, \ldots, \hat{p}_t) = l_L$ are the log-likelihood functions of the "reduced" and the "large model" evaluated at their maxima ($\hat{p}$ and $(\hat{p}_1, \ldots, \hat{p}_t)$), respectively.

Review:

The χ^2 distribution

- $X_1, \ldots, X_n$ iid
  $X_1 \sim N(0,1)$
- $Z = X_1^2 + \cdots + X_n^2$ has a known distribution
  called the chi-square distribution with n degrees of freedom
- Notation $Z \sim \chi^2(n)$
- $E(Z) = n$

Review:

The general idea of the likelihood ratio test

- It can be shown that, if $p_1 = \cdots = p_t$, then Λ is approximately chi-square distributed (for values of n large enough).
- The number of degrees of freedom d of the referred chi-square distribution is given by the difference between the number of parameters of the "larger model" , d_l, minus the number of parameters of the "reduced model", d_r, i.e. $d = d_l - d_r$.
- The quantity Λ can be used to test the null hypothesis

  $$H_0 : p_1 = p_2 = \cdots = p_t,$$

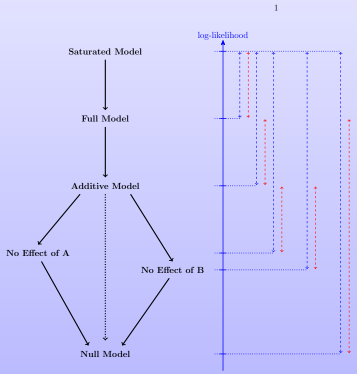
  at a level of significance α, by using the rule

  $$\text{" Reject } H_0 \text{ when } \Lambda \geq \chi^2_d(1-\alpha) \text{ ".}$$

- Here $\chi^2_d(1-\alpha)$ is the $(1-\alpha)$- quantil of a chi-square distribution with d degrees of freedom.

Review:

Possible two-way classification models



- The Poisson distribution is often used to model number of events such as number of accidents, number of mutations in a fragment of DNA, number of worms in a portion of soil, etc.
- This distribution was first used by Siméon-Denis Poisson
  Poisson, S.D., 1838. *Recherches sur la probabilité des jugements en matières criminelles et matière civile*
  (Study on the Probability of Judgments in Criminal and Civil Matters)
  to study the number of occurrences of an event during a time-interval of a given length, specifically the number of criminal and civil judgments
- The Poisson distribution takes positive integer values (*i.e.* 0,1,2, . . .) and depends on a single parameter, called the *intensity parameter* and usually denoted by λ

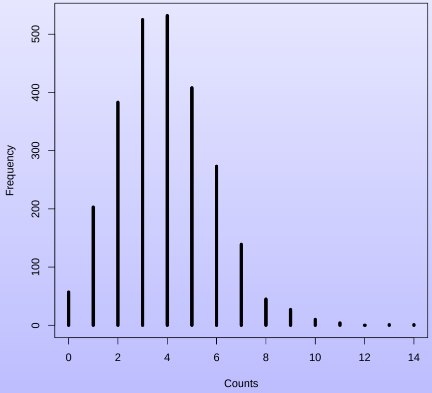
A classical example - Counts of alpha-particles

- Frequency of counts of alpha-particles emitted by the radioactive decay of a source of polonium, registered in time-intervals of 72 seconds .

Counts:	0	1	2	3	4	5	6	7
Frequency:	57	203	383	525	532	408	273	139
Counts:	8	9	10	11	12	13	14	+ 15
Frequency:	45	27	10	4	0	1	1	0

- Rutherford, E. and Geiger, M. (1910). The probability variations in the distribution of alpha-particles. *Philosophical Magazine*, series 6, **20**, 698-704.

A classical example - Counts of alpha-particles



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

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

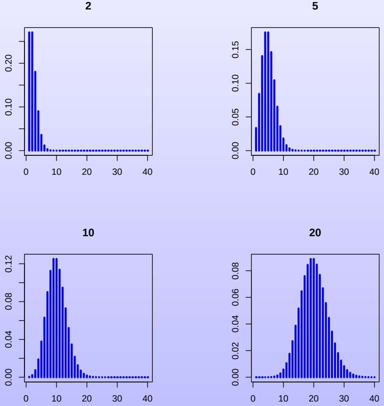
  for $y = 0, 1, 2, \ldots$.
  Here $y! = y \cdot (y-1) \cdot \cdots \cdot 1$ and $0! = 1$.

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

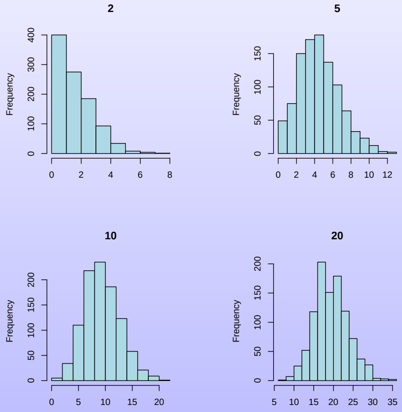
- Notation: $Y \sim Po(\lambda)$

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

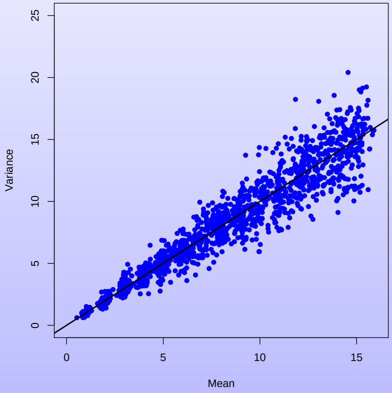
Probability function of the Poisson distribution



Simulated 1000 Poisson random variables



Simulated Poisson with different means: mean X variance

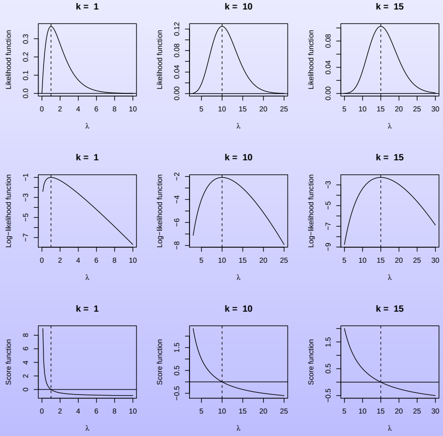


Parameter Estimation for a simple Poisson model

The statistical model

- $Y_1, \dots Y_n$ iid $Y_1 \sim Po(\lambda)$
- The likelihood function for observations $y_1, \dots y_n$ is
  $L(\lambda) = \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} \cdot \dots \cdot \frac{e^{-\lambda}\lambda^{y_n}}{y_n!}$
- The log-likelihood is
  $l(\lambda) = -n\lambda + \log(\lambda)\sum_i y_i + K$
  K is a constant depending on $y_1, \dots, y_n$ but **not** on λ
- The score function is
  $S(\lambda) = \frac{\partial}{\partial\lambda} l(\lambda) = -n + \frac{1}{\lambda}\sum_{i=1}^n y_i$
- Equating the score function to zero yields
  $\frac{1}{\hat{\lambda}}\sum_{i=1}^n y_i = n$
  which has solution $\hat{\lambda} = \frac{\sum_{i=1}^n y_i}{n}$
- The sample mean is the maximum likelihood estimate for λ

Likelihood quantities for simple Poisson models



k = observed total number of counts.

A classical example - Counts of alpha-particles

- Frequency of counts of alpha-particles emitted by the radioactive decay of a source of polonium, registered in time-intervals of 72 seconds

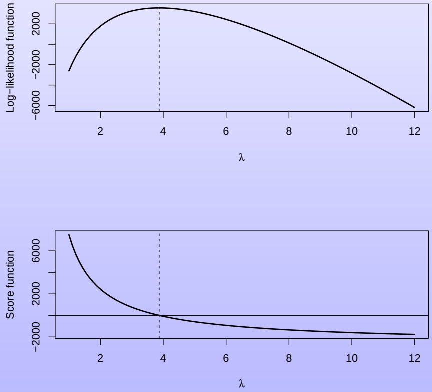
Counts:	0	1	2	3	4	5	6	7
Frequency:	57	203	383	525	532	408	273	139
Counts:	8	9	10	11	12	13	14	+ 15
Frequency:	45	27	10	4	0	1	1	0

Rutherford, E. and Geiger, M. (1910).

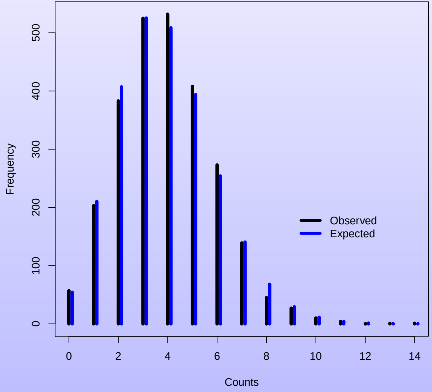
- Mean of counts: 3.87

  Variance of counts: 3.74
- The maximum likelihood estimate of λ is 3.87

Likelihood quantities for a Poisson models for the Rutherford-Geiger data



A classical example - Counts of alpha-particles



Histogram of the counts per interval and the expected number of counts for a Poisson distributed random variable.

A classical example - Counts of alpha-particles

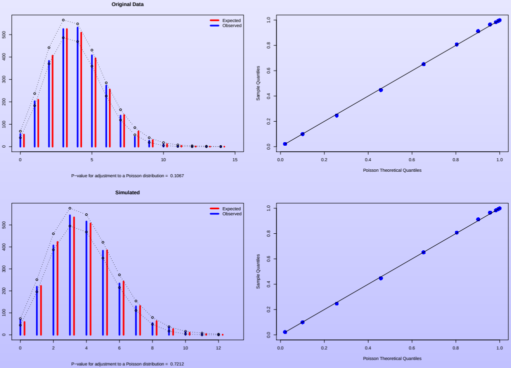
- Another way check the adequacy of a Poisson distribution here is to to plot the observed quantiles against the expected quantiles under the hypothesis of Poisson distribution
- This is called the Poisson QQ-plot.

```
qqpois <- function(x, lambda, main=" ",){
  ox <- x[order(x)]
  emp <- ecdf(ox)(ox)
  teor <- ppois(ox, lambda=lambda)
  plot(teor,emp, xlab="Poisson Theoretical Quantiles", ylab="Sample Quantiles",
       main=main,ylim=c(0,1), cex=1.5, pch=19, col=’blue’)
    lines(c(min(emp),max(emp)), c(min(emp),max(emp)), col="black", lwd=2)
}
```

A classical example - Counts of alpha-particles

```
# Reconstructing the individual counts from the frequency of counts
Obs <- c(57 , 203 , 383 , 525 , 532 , 408 , 273 , 139 ,45 ,  27 , 10 , 4 ,  0 ,  1 ,1)
Y <- rep(0, Obs[1])
for(j in 2:15){
  Aux <- rep(j-1, Obs[j])
  Y <- c(Y, Aux)
}
mean(indvidual.counts)  # 3.877778
var(indvidual.counts)   # 3.743967
# Drawing the QQ-plot
qqpois(indvidual.counts, lambda=lambda)
```

Rutherford Geiger Data: Checking the Adherence to the Poisson Distribution



A classical example - Counts of alpha-particles

```
P-value =  0.1005
 Value  Exp. Freq. Obs. Freq.   Lower   Upper
     0  54.0185616         57  39.000  69.000
     1 209.4719779        203 185.000 237.000
     2 406.1428906        383 371.000 444.000
     3 524.9772921        525 482.975 564.025
     4 508.9363195        532 471.000 549.000
     5 394.7083902        408 360.000 431.000
     6 255.0985708        273 225.000 282.000
     7 141.3165099        139 119.975 166.000
     8  68.4992528         45  53.000  85.000
     9  29.5138756         27  19.000  40.000
    10  11.4448251         10   5.000  19.000
    11   4.0345899          4   1.000   8.000
    12   1.3037702          2   0.000   4.000
    13   0.3889024          1   0.000   2.000
    14   0.1077198          1   0.000   1.000
```

Classical examples of use of the Poisson distribution:

- The number of α particles emitted from a radioactive substance in a fixed time interval
(Rutherford, Geiger and Bateman, 1910).
- The number of yeast cells per cube in an hemacytometer
(Student, 1907, *i.e.* William Gosset)
- The number of telephone calls arriving in a telephone central per unit time
(Erlang, 1909, in Danish!)
- The number of deaths by horse kicks in the Prussian army!

Two further examples

- We will discuss in the next two lectures two examples 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)

The Poisson regression

Example: Penicillium in soil

- *Penicillium verrucosum* is a fungus that infects grains and produces a toxin.
- *P. verrucosum* survives in soil and subsequently infests grains.
- We want to determine the abundance of **C**olony **F**orming **U**nits (CFU) in soil samples.

  (i.e. how many CFU / g soil).

The Poisson regression

Penicillium in soil

The Poisson regression

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
- Better method:
  Use several amounts of soil and assume that the expected number of CFU is proportional to the amount of soil added

The Poisson regression

Example: Penicillium in soil

The Poisson regression

Example: Penicillium in soil

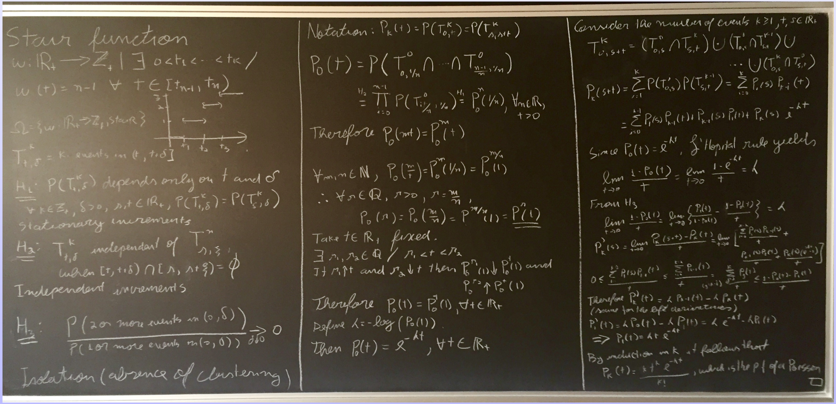
The Poisson regression

Example: Penicillium in soil

- The probability distribution of the number of colonies per Petri dish can be deduced (under some reasonable assumptions)!
- 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 in an optional section of the notes, involves a proper formulation of the problem as a stochastic process and the solution of a differential equation)

Deducing the Poisson Distribution



Stair function

$w: \mathbb{R}_+ \longrightarrow \mathbb{Z}_+ \mid \exists\, 0<t_1<\cdots<t_k \,/$

$w(t) = n-1 \;\forall\, t \in [t_{n-1}, t_n)$

$\Omega = \{w: \mathbb{R}_+ \to \mathbb{Z}_+, \text{stair}\}$

$T^k_{t,\delta}$ = k events in $(t, t+\delta]$

H1: $P(T^k_{t,\delta})$ depends only on t and δ

$\forall k \in \mathbb{Z}_+, \delta > 0, s,t \in \mathbb{R}_+, P(T^k_{t,\delta}) = P(T^k_{s,\delta})$

Stationary increments

H2: $T^k_{t,\delta}$ independant of $T^n_{s,\xi}$,

when $[t, t+\delta) \cap [s, s+\xi) = \phi$

Independent increments

H3: $\dfrac{P(\text{2 or more events in } (0,\delta))}{P(\text{1 or more events in } (0,\delta))} \xrightarrow[\delta\downarrow 0]{} 0$

Isolation (absence of clustering)

Notation: $P_k(t) = P(T^k_{0,t}) = P(T^k_{s,s+t})$

$P_0(t) = P(T^0_{0,1/n} \cap \cdots \cap T^0_{\frac{n-1}{n},1/n})$

$\overset{H_2}{=} \prod_{i=0}^{n-1} P(T^0_{i/n,1/n}) \overset{H_1}{=} P_0^n(1/n), \; \forall n \in \mathbb{R}, t>0$

Therefore $P_0(mt) = P_0^m(t)$

$\forall m,n \in \mathbb{N}, \; P_0(\tfrac{m}{t}) = P_0^m(1/n) = P_0^{m/n}(1)$

$\therefore \forall r \in \mathbb{Q}, r>0, r = \frac{m}{n},$

$P_0(r) = P_0(\tfrac{m}{n}) = P^{m/n}(1) = \underline{\underline{P^r(1)}}$

Take $t \in \mathbb{R}_+$ fixed.

$\exists\, r_1, r_2 \in \mathbb{Q} \,/\, r_1 < t < r_2$

If $r_1 \uparrow t$ and $r_2 \downarrow t$ then $P_0^{r_1}(1) \downarrow P_0^t(1)$ and $P_0^{r_2} \uparrow P_0^t(1)$

Therefore $P_0(t) = P_0^t(1), \forall t \in \mathbb{R}_+$

Define $\lambda = -\log(P_0(1))$.

Then $P_0(t) = e^{-\lambda t}, \; \forall t \in \mathbb{R}_+$

Consider the number of events $k \geq 1, t, s \in \mathbb{R}_+$

$T^k_{0,s+t} = (T^0_{0,s} \cap T^k_{s,t}) \dot\cup (T^1_{0,s} \cap T^{k-1}_{s,t}) \dot\cup \cdots \dot\cup (T^k_{0,s} \cap T^0_{s,t})$

$P_k(s+t) = \sum_{i=1}^{k} P(T^i_{0,s}) P(T^{k-i}_{s,t}) = \sum_{i=0}^{k} P_i(s) P_{k-i}(t)$

$= \sum_{i=0}^{k-1} P_i(s) P_{k-i}(t) + P_{k-1}(s) P_1(t) + P_k(s)\, e^{-\lambda t}$

Since $P_0(t) = e^{-\lambda t}$, L'Hopital rule yields

$\lim_{t\to 0} \dfrac{1-P_0(t)}{t} = \lim_{t\to 0} \dfrac{1-e^{-\lambda t}}{t} = \lambda$

From H3

$\lim_{t\to 0} \dfrac{1-P_0(t)}{t} = \lim_{t\to 0} \left\{ \dfrac{P_1(t)}{1-P_0(t)} \cdot \dfrac{1-P_0(t)}{t} \right\} = \lambda$

$P_k'(s) = \lim_{t\to 0} \dfrac{P_k(s+t) - P_k(t)}{t} = \lim_{t\to 0} \left[\dfrac{\sum_{i=0}^{k-2} P_i(s) P_{k-i}(t)}{t} + \dfrac{P_{k-1}(s) P_1(t)}{t} + \dfrac{P_k(s)(e^{-\lambda t}-1)}{t} \right]$

$0 \leq \dfrac{\sum_{i=0}^{k-2} P_i(s) P_{k-i}(t)}{t} \leq \dfrac{\sum_{i=0}^{k-2} P_{k-i}(t)}{t} \underset{(j=k-i)}{=} \dfrac{\sum_{j=2}^{k} P_j(t)}{t} \leq \dfrac{1-P_0(t)-P_1(t)}{t}$

Therefore $P_k'(t) = \lambda P_{k-1}(t) - \lambda P_k(t)$

(same for the left derivatives)

$P_1'(t) = \lambda P_0(t) - \lambda P_1(t) = \lambda e^{-\lambda t} - \lambda P_1(t)$

$\Rightarrow P_1(t) = \lambda t\, e^{-\lambda t}$

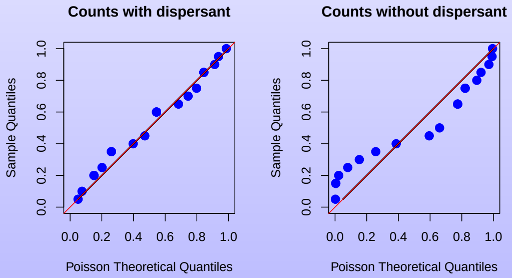
By induction on k it follows that

$P_k(t) = \dfrac{\lambda^k t^k e^{-\lambda t}}{k!}$, which is the pf of a Poisson

□

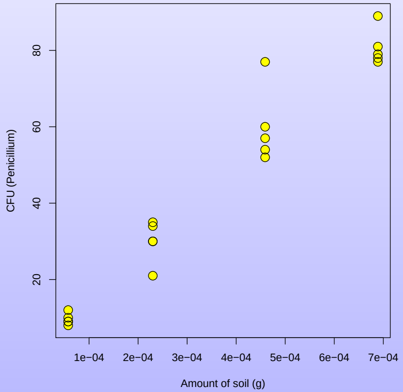
Penicillium in soil

Preliminary experiment without and with dispersant



The Poisson regression

Penicillium in soil



The Poisson regression

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
- If $Y_{g,d} \sim Po(\lambda)$ then the expected number of CFU in this Petri dish is $E(Y_{g,d}) = \lambda$.

- Saturated model:
  $Y_{g,d} \sim Po(\lambda_{g,d})$
- Platting method model (linear0):
  $Y_{g,d} \sim Po(\lambda_{g,d})$
  $E(Y_{g,d}) = \lambda_{g,d} = \beta g$
- Here β is a parameter.
  Interpretation of β: Number of CFU per gram soil! (why?)

The Poisson regression

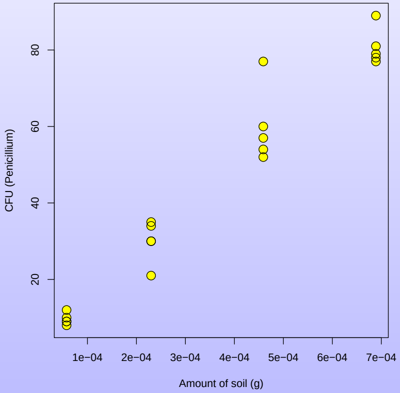
Penicillium in soil

- Saturated model:
  $Y_{g,d} \sim Po(\lambda_{g,d})$
- Free curve model:
  $$Y_{g,d} \sim Po(\lambda_{g,d}) \; E(Y_{g,d}) = \lambda_{g,d} = \beta_g = \begin{cases} \beta_{0.00000574} \\ \beta_{0.00023000} \\ \beta_{0.00045900} \\ \beta_{0.00068900} \end{cases}$$
  (i.e. one expectation for each g)
- Linear model:
  $Y_{g,d} \sim Po(\lambda_{g,d})$
  $E(Y_{g,d}) = \lambda_{g,d} = \alpha + \beta g$
  Interpretation of α and β?
- Platting method model (linear0):
  $Y_{g,d} \sim Po(\lambda_{g,d})$
  $E(Y_{g,d}) = \lambda_{g,d} = \beta g$

Penicillium in soil Calculations in R

```
> ################################################################################

> # Poisson regression model                                                     #

> ################################################################################


> attach(data.fungi)


> str(data.fungi)

’data.frame’:        20 obs. of  3 variables:

 $ Penicillium: int  9 8 9 12 10 34 30 35 30 21 ...

 $ other      : int  3 2 3 2 3 13 17 8 13 10 ...

 $ gsoil      : num  5.74e-05 5.74e-05 5.74e-05 5.74e-05 5.74e-05 2.30e-04  ...
```

Penicillium in soil



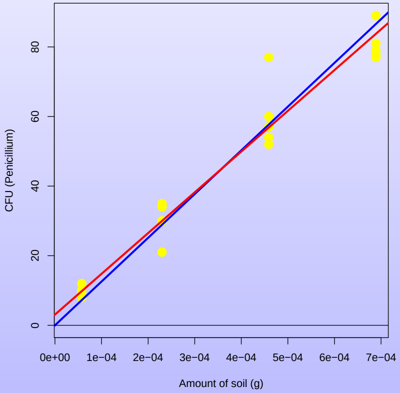
Penicillium in soil Test of homogeneity, estimating the abundance

```
> free.curve <- glm(Penicillium ~ factor(gsoil) , family=poisson(link="log") )
> # Testing homogeneity
> deviance(free.curve)
[1] 12.68678
> length(gsoil) # 20 observations  and 4 parameters in the free curve model
[1] 20
> pchisq(deviance(free.curve), df=16, lower.tail=F)
[1] 0.6955064
> # Fitting a linear model through the origin
> linear0 <- glm(Penicillium ~ 0 + gsoil , family=poisson(link="identity") )
> coef(linear0)
   gsoil
125679.3
```

Penicillium in soil Fitting a linear model and comparing with a plating model

```
> # Fitting a linear model (not necessarily through the origin)
> linear <- glm(Penicillium ~ gsoil , family=poisson(link="identity") )
> coef(linear)
 (Intercept)        gsoil
3.074457e+00 1.171117e+05
> coef(linear0)
   gsoil
125679.3
```

Penicillium in soil Comparing a linear model and with a plating model



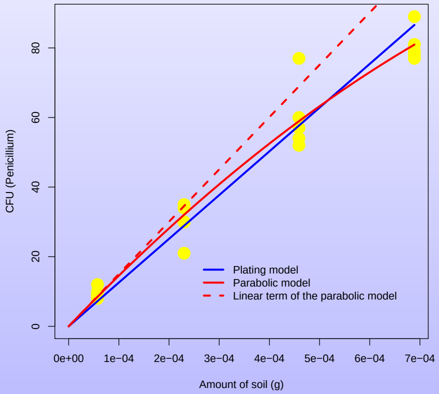
```
> # Testing linearity
> anova(linear, free.curve, test="Chisq")
Analysis of Deviance Table
Model 1: Penicillium ~ gsoil
Model 2: Penicillium ~ factor(gsoil)
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        18    14.1064
2        16    12.6868  2   1.4196    0.4917
> # Testing whether the regression line crosses the origin
> anova(linear0, linear, test="Chisq")
Analysis of Deviance Table
Model 1: Penicillium ~ 0 + gsoil
Model 2: Penicillium ~ gsoil
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        19    18.9024
2        18    14.1064  1   4.7960    0.0285
```

Modelling competition

- Platting method model (linear0):
  $Y_{g,d} \sim Po(\lambda_{g,d})$
  $E(Y_{g,d}) = \lambda_{g,d} = \beta g$
- A larger model: (parabolic)
  $Y_{g,d} \sim Po(\lambda_{g,d})$
  $E(Y_{g,d}) = \lambda_{g,d} = \beta g + \gamma g^2$
- The last term (γg^2) allow to represent competition
  (if $\gamma < 0$)
  Could have incorporated also higher order polynomials.
- Interpretation of β changes in the large model!
  What is the change in the interpretation?

```
> #  Fitting a parabolic regression
> gsoil2 <- gsoil*gsoil
> parabola <- glm(Penicillium ~ 0 + gsoil + gsoil2 , family=poisson(link="identity") )
> anova(linear0, parabola, test="Chisq")
Analysis of Deviance Table
Model 1: Penicillium ~ 0 + gsoil
Model 2: Penicillium ~ 0 + gsoil + gsoil2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        19    18.9024
2        18    14.1721  1   4.7302    0.0296
> coef(parabola)
      gsoil      gsoil2
   150264.1 -47585250.7
> coef(linear0)
   gsoil
125679.3
```

Modelling competition



Do we have competition or contamination?

Idea: Compare the parabolic and the linear model to a common reference, the free-curve model

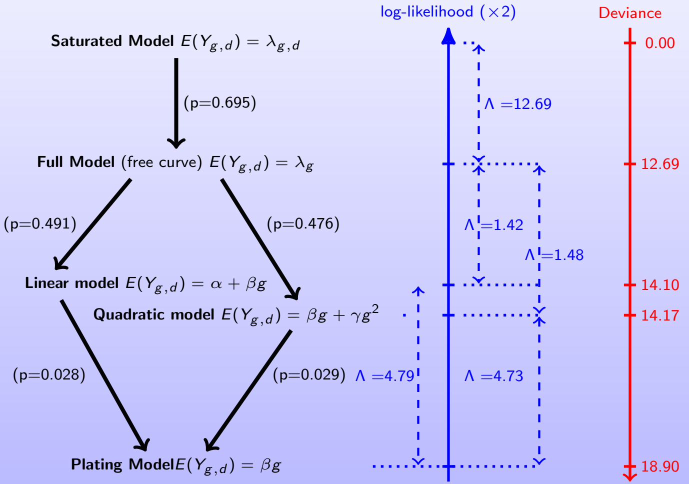
```
> free.curve <- glm(Penicillium ˜ factor(gsoil) , family=poisson(link="log") )
> parabola <- glm(Penicillium ˜ 0 + gsoil + gsoil2 , family=poisson(link="identity") )
> linear <- glm(Penicillium ˜ gsoil , family=poisson(link="identity") )
> anova(parabola, free.curve, test="Chisq")
Analysis of Deviance Table
Model 1: Penicillium ˜ 0 + gsoil + gsoil2
Model 2: Penicillium ˜ factor(gsoil)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        18     14.172
2        16     12.687  2   1.4853   0.4758
```

Do we have competition or contamination?

Idea: Compare the parabolic and the linear model to a common reference, the free-curve model

```
> free.curve <- glm(Penicillium ~ factor(gsoil) , family=poisson(link="log") )
> parabola <- glm(Penicillium ~ 0 + gsoil + gsoil2 , family=poisson(link="identity") )
> linear <- glm(Penicillium ~ gsoil , family=poisson(link="identity") )
> anova(linear, free.curve, test="Chisq")
Analysis of Deviance Table
Model 1: Penicillium ~ gsoil
Model 2: Penicillium ~ factor(gsoil)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        18     14.106
2        16     12.687  2   1.4196   0.4917
```

Summing up



Saturated Model $E(Y_{g,d}) = \lambda_{g,d}$

(p=0.695)

Full Model (free curve) $E(Y_{g,d}) = \lambda_g$

(p=0.491)

(p=0.476)

Linear model $E(Y_{g,d}) = \alpha + \beta g$

Quadratic model $E(Y_{g,d}) = \beta g + \gamma g^2$

(p=0.028)

(p=0.029)

Plating Model$E(Y_{g,d}) = \beta g$

log-likelihood ($\times 2$)

Λ =12.69

Λ =1.42

Λ =1.48

Λ =4.79

Λ =4.73

Deviance

0.00

12.69

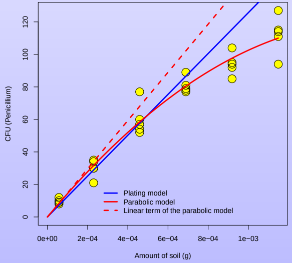
14.10

14.17

18.90

Do we have competition or contamination?

- Cannot decide, on the basis of this data, whether we have competition (parabolic model) or contamination (linear model)
- To solve this in definition we extended the data, obtained by adding observations with more soil



Do we have competition or contamination? Analysis of the **extended data**.

Testing homogeneity

```
> free.curve <- glm(Penicillium ~ factor(gsoil) , family=poisson(link="log") )
> parabola <- glm(Penicillium ~ 0 + gsoil + gsoil2 , family=poisson(link="identity") )
> linear <- glm(Penicillium ~ gsoil , family=poisson(link="identity") )
> linear0 <- glm(Penicillium ~ 0 + gsoil , family=poisson(link="identity") )
> deviance(free.curve)
[1] 19.7676
> pchisq(deviance(free.curve), df=24, lower.tail=F)
[1] 0.7099128
```

Do we have competition or contamination? Analysis of the **extended data**.

Testing adequacy of the quadratic model (competition)

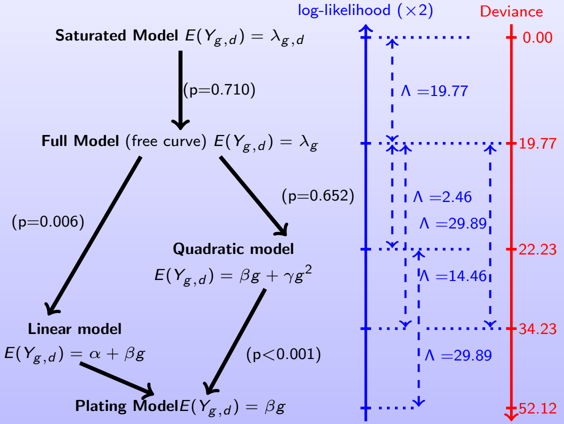
```
> free.curve <- glm(Penicillium ~ factor(gsoil) , family=poisson(link="log") )
> parabola <- glm(Penicillium ~ 0 + gsoil + gsoil2 , family=poisson(link="identity") )
> linear <- glm(Penicillium ~ gsoil , family=poisson(link="identity") )
> anova(parabola, free.curve, test="Chisq")
Analysis of Deviance Table
Model 1: Penicillium ~ 0 + gsoil + gsoil2
Model 2: Penicillium ~ factor(gsoil)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        28     22.226
2        24     19.768  4   2.4584   0.6521
```

Do we have competition or contamination? Analysis of the **extended data**.

Testing adequacy of the linear model (contamination)

```
> free.curve <- glm(Penicillium ~ factor(gsoil) , family=poisson(link="log") )
> parabola <- glm(Penicillium ~ 0 + gsoil + gsoil2 , family=poisson(link="identity") )
> linear <- glm(Penicillium ~ gsoil , family=poisson(link="identity") )
> anova(linear, free.curve, test="Chisq")
Analysis of Deviance Table
Model 1: Penicillium ~ gsoil
Model 2: Penicillium ~ factor(gsoil)
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         28     34.233
2         24     19.768  4   14.466 0.005948 **
```

Do we have competition or contamination? Analysis of the **extended data**



Saturated Model $E(Y_{g,d}) = \lambda_{g,d}$

(p=0.710)

Full Model (free curve) $E(Y_{g,d}) = \lambda_g$

(p=0.006)

(p=0.652)

Quadratic model $E(Y_{g,d}) = \beta g + \gamma g^2$

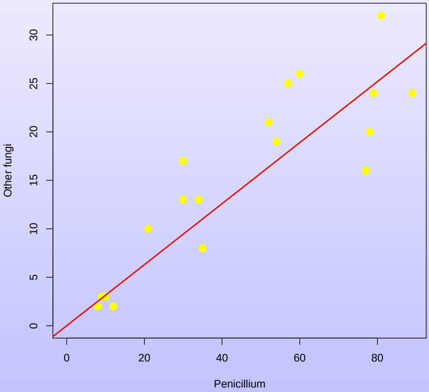
Linear model $E(Y_{g,d}) = \alpha + \beta g$

(p<0.001)

Plating Model $E(Y_{g,d}) = \beta g$

Modelling the competition

- There were also other fungi in the soil,
  recorded as the number of CFU of other fungi
- Natural question:
  How much intra and inter-specific competition explain the observed effect of competition on Penicillium?
- We use a binomial model to verify whether the competition is intra- or inter-specific
- We use the abundance of other fungi as an explanatory variable in the parabolic Poisson regression to test inter-specific competition


Other fungi
0
5
10
15
20
25
30
0
20
40
60
80
Penicillium

Modelling the competition: Fitting a binomial model for the probability of being a Penicillium CFU

```
> resp <- cbind(Penicillium, other)
> binom <- glm(resp ~ factor(gsoil), family=binomial)
> anova(binom, test="Chisq")
              Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                             19    19.0024
factor(gsoil)  3   4.5091        16    14.4933    0.2115
```

- The proportion of Penicillium does not increase with the amount of soil added
- Suggests that the competition is intra-specific.

Modelling the competition

```
> compet.corr <- glm(Penicillium ~ 0 + other + gsoil + gsoil2  , family=poisson(link="identity") )
> anova(compet.corr, test="Chisq")
Analysis of Deviance Table
Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      20        Inf
other   1      Inf        19         79 < 2.2e-16 ***
gsoil   1       60        18         18 8.623e-15 ***
gsoil2  1        5        17         14    0.0332 *
```

- The there is still a significant competition effect (gsoil2) even after correcting for the increasing presence of other fungi.

Modelling the competition

```
> anova(parabola,  compet.corr, test="Chisq")
Analysis of Deviance Table

Model 1: Penicillium ~ 0 + gsoil + gsoil2
Model 2: Penicillium ~ 0 + other + gsoil + gsoil2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        18    14.1721
2        17    13.8005  1   0.3716    0.5421
```

- There is no significant effect of competition of Penicillium by other species when the possible effect of competition of Penicillium by Penicillium is accounted for.

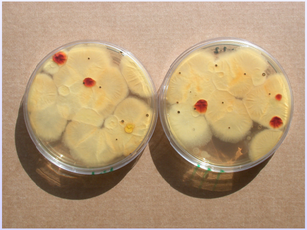
Modelling the competition

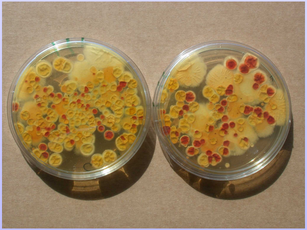
```
> summary(compet.corr)
Call:
glm(formula = Penicillium ~ 0 + other + gsoil + gsoil2, family = poisson(link = "identity"))
...
Coefficients:
         Estimate Std. Error z value Pr(>|z|)
other  -2.683e-01  4.429e-01  -0.606   0.5447
gsoil   1.664e+05  2.983e+04   5.578 2.44e-08 ***
gsoil2 -5.727e+07  2.793e+07  -2.051   0.0403 *
...
```

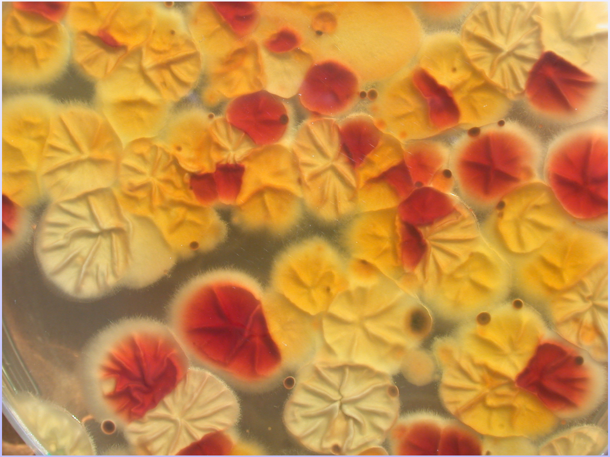
Modelling the competition

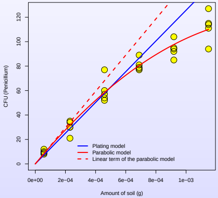
```
> no.compet <- glm(Penicillium ~ 0 + other + gsoil    ,
+                   family=poisson(link="identity") )
> anova(no.compet, compet.corr, test="Chisq")
Analysis of Deviance Table

Model 1: Penicillium ~ 0 + other + gsoil
Model 2: Penicillium ~ 0 + other + gsoil + gsoil2
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        18    18.3359
2        17    13.8005  1   4.5354    0.0332 *
```

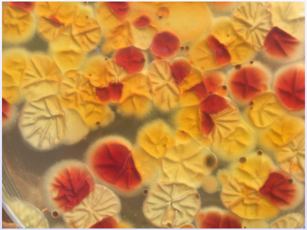
- Competition cannot be explained by the presence of other fungi

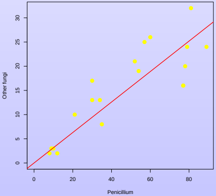












In conclusion,
Penicillium verrucosum is not like *Homo sapiens sapiens*,
when there is lack of resources they do not kill the other species!

Summary of the day

What we have covered today:

- The Poisson distribution
- Some techniques for verifying whether a variable is Poisson distributed Poisson QQ-plots, tests
- Inference under Poisson models (MLE, LRT, etc)
- A set of basic assumptions for deducing that a variable is Poisson distributed
- Example of linear and non-linear regression using a Poisson model
- Several examples of the interface between a biological discussion and mathematical and statistical modelling