Code

Big Ideas for Small Data : Some Bayesian Ideas learnt from the past

Eric Parent

the 13th of May, 2019

What do statisticians do when sufficient data is available?

- …the great potential of using data to help us understand the world and make better judgements. This is what statistical science is all about.

(David Spiegelhalter, The Art of Statistics. Learning from Data)

- Big worlds & Small worlds…

(Ben Lambert, A Student's Guide to Bayesian Statistics)

- A science of interpretation?

Statistics/probability

Cooking/chemistry

Psychology

What statisticians do when few data is available?

- They do not know much…
    - Some hide their helplessness under the carpet…
    - The language of probability is taylored to be honest : Probabilistic assessments of credibility! Especially convenient for Bayesians.
- They search for additional *sources of information*…
    - Assuming hypotheses as a *model based* way to (artificially?) reduce uncertainty.
    - Off the shelf auxiliary *theories*:
        - symmetry & principle of (un)sufficient reason,
        - entropy,
        - phenomenological mockery
        - recourse to asymptotic behavior

Trying to see farther by standing on former giants’shoulders:

Much of the following material has been stolen from:

- Tribus: Rational Descriptions, Decisions & Designs, pages 396–404 (https://www.elsevier.com/books/rational-descriptions-decisions-and-designs/tribus/978-0-08-006393-5)


- Maranzano & Krzyzstofowicz: Bayesian Reanalysis of the Challenger O-Ring Data, Risk Analysis, Vol. 28, No. 4, 2008

Part 1:
Decision under uncertainty with very few information

The widget factory of Jaynes

You are the newly appointed director of color styling in a widget factory.

Suppose the factory produces 200 widgets each day, which must all be painted either red, blue or green.

The company has issued an advertisement which reads, “Delivery in 24 hours or double your money back”.

Mise en situation décisionnelle:
Vous fabriquez···.

Maison Rouge

Maison Bleue

Maison Verte

2OO objets fabriqués par jour
Un seul modèle par jour (couleur)

First state of knowledge

Since you are new on the job, you will try to find out as much as you can before making a decision.

Suppose you look into the storeroom and find the following information:

First state of knowledge (cont’d)

Mise en situation décisionnelle:
Première information: *niveau du stock*

	ETAT DES STOCKS	
ROUGE	BLEUE	VERTE
100	150	50

2OO objets produits par jour
Un seul modèle

Quelle couleur appliqueriez-vous aujourd'hui ?

Second state of knowledge

Of course, this is not very good information on which to base a decision which might cost you a considerable amount of money. Suppose, therefore, that you gathered more information and found that the average daily orders were as indicated in the following table:

Second state of knowledge (cont'd)

Mise en situation décisionnelle:
Seconde information: *Demande journalière*
(*en nombre total de maisonnettes*)

	ROUGE	BLEUE	VERTE
STOCKS	100	150	50
COMMANDE QUOTIDIENNE MOYENNE	50	100	10

2OO objets par jour
Un seul modèle

Quelle couleur appliqueriez-vous aujourd'hui ?

Third state of knowledge

Even with this information, you would not like to make the decision, you would certainly want more information. Suppose you were to learn something about the size of the individual orders, as indicated in the table below:

Third state of knowledge (cont'd)

Mise en situation décisionnelle:
Troisième info: *Nombre moyen d'objets par commande journalière*

	ROUGE	BLEUE	VERTE
STOCKS	100	150	50
COMMANDE QUOTIDIENNE MOYENNE	50	100	10
TAILLE MOYENNE DE CHAQUE COMMANDE	75	10	20

2OO objets produits par jour
Un seul modèle

Quelle couleur appliqueriez-vous aujourd'hui?

Fourth state of knowledge

Using this third stage of knowledge, generally 80 per cent of the audience will agree with the choice given by the vote.

Yet many people will not be able to give a cogent argument supporting their view.

At this point that most engineers will concede the usefulness of a mathematical analysis.

However, just to make the point , suppose you receive a long distance telephone call which advises you that an order has just been received for 40 green widgets. In the fourth state of knowledge, the information appears as follows:

Fourth state of knowledge (cont'd)

Mise en situation décisionnelle:
Quatrième info: *On vient de recevoir une commande!*

STOCKS	100	150	50
COMMANDE QUOTIDIENNE MOYENNE	50	100	10
TAILLE MOYENNE DE CHAQUE COMMANDE	75	10	20
COMMANDE DÉJÀ REÇUE	-	-	40

2OO objets produits par jour
Un seul modèle

Quelle couleur appliqueriez-vous aujourd'hui?

The need for modeling (state of knowledge 1)

Hide

Hide

```
rampe=function(n,s){ # n nombre de commandes, s stock
  res= (n-s)*(n>=s)  # perte si les commandes dépassent le stock
}
```

$$L(n_r, n_b, n_v, d = i \,|Info) = \sum_{j=r,b,v} \sum_{n_j=0}^{\infty} r(n_j - 200\delta_{ij} - s_j) p(n_j \,|Info)$$

First state of knowledge : symmetry!

$$p(N_r = n \,|I_1) = p(N_b = n \,|I_1) = p(N_v = n \,|I_1)$$

then $\sum_{j=r,b,v} r(n_j - 200\delta_{ij} - s_j)$ is minimal for $i = v$

Hide

Hide

```
stock=c(100,150,50)
U=0*stock
fabrication=0*stock
for (couleur in 1:3) {fab=fabrication
                      fab[couleur]=200
                      U[couleur]=sum(rampe(n=200,stock+fab))
}
head(U)
```

```
[1] 200 250 150
```

MaxEnt formalism : Implementation procedure

1. Define the permitted events
2. Collect available information according to these contingencies
3. Find the probability function maximizing entropy $S = -\sum_{i=1}^{M} p_i log(p_i)$ while meeting the constraints imposed by the information available. Generally they are of the form: $\bar{g_k} = \sum_{i=1}^{M} p_i \times g_r(i)$ for some functions $\{g_k(\cdot)\}_{k=1:r}$
4. Then update this distribution using Bayes' formula each time new information becomes available.

MaxEnt formalism : What to do?

Solve:

$$\begin{aligned} Max \quad & - \sum_{i=1}^{M} p_i log(p_i) \\ \sum_{i=1}^{M} p_i \quad & = \quad 1 \\ \sum_{i=1}^{M} p_i \times g_1(i) \quad & = \quad \bar{g_1} \\ \cdots & \\ \sum_{i=1}^{M} p_i \times g_r(i) \quad & = \quad \bar{g_r} \end{aligned}$$

Differentiate:

$$L = \sum_{i=1}^{M} p_i log(p_i) + (\lambda_0 - 1)(\sum_{i=1}^{M} p_i) + \sum_{k=1}^{r} (\lambda_k) \left(\sum_{i=1}^{M} p_i \times g_k(i) \right)$$

and get $p_i = exp(-\lambda_0 - \sum_{k=1}^{r} (\lambda_k g_k(i)))$ with:

$$\lambda_0 = log(\sum_{i=1}^{M} exp(- \sum_{k=1}^{r} (\lambda_k g_k(i))))$$

$$-\frac{\partial \lambda_0}{\partial \lambda_k} = \bar{g_k} = E(g_k(X))$$

$$\frac{\partial^2 \lambda_0}{\partial \lambda_j \partial \lambda_k} = \mathrm{cov}(g_j(X) g_k(X))$$

$$S^* = {\lambda_0}^* + \sum_{k=1}^{r} {\lambda_k}^* \bar{g_k}$$

MaxEnt formalism: Application

Exponential distribution $(\bar{g}_1) = \bar{X} = \mu$ with $i \in \mathbb{N}$, then: $\lambda_0 = log(1+\mu)$ by computing:

$$-\frac{\partial \lambda_0}{\partial \lambda_1} = \mu$$

and get :

$$p_i = \frac{1}{1+\mu}\left(\frac{\mu}{1+\mu}\right)^i$$

$$S^* = (1+\mu)log(1+\mu) - \mu log(\mu)$$

Second state of knowledge (MaxEnt)

From MAXENT principle we get:

$$p(n_r \,|I_2) = \frac{1}{51}\left(\frac{50}{51}\right)^{n_r}$$
$$p(n_b \,|I_2) = \frac{1}{101}\left(\frac{100}{101}\right)^{n_b}$$
$$p(n_v \,|I_2) = \frac{1}{11}\left(\frac{10}{11}\right)^{n_v}$$

so that:

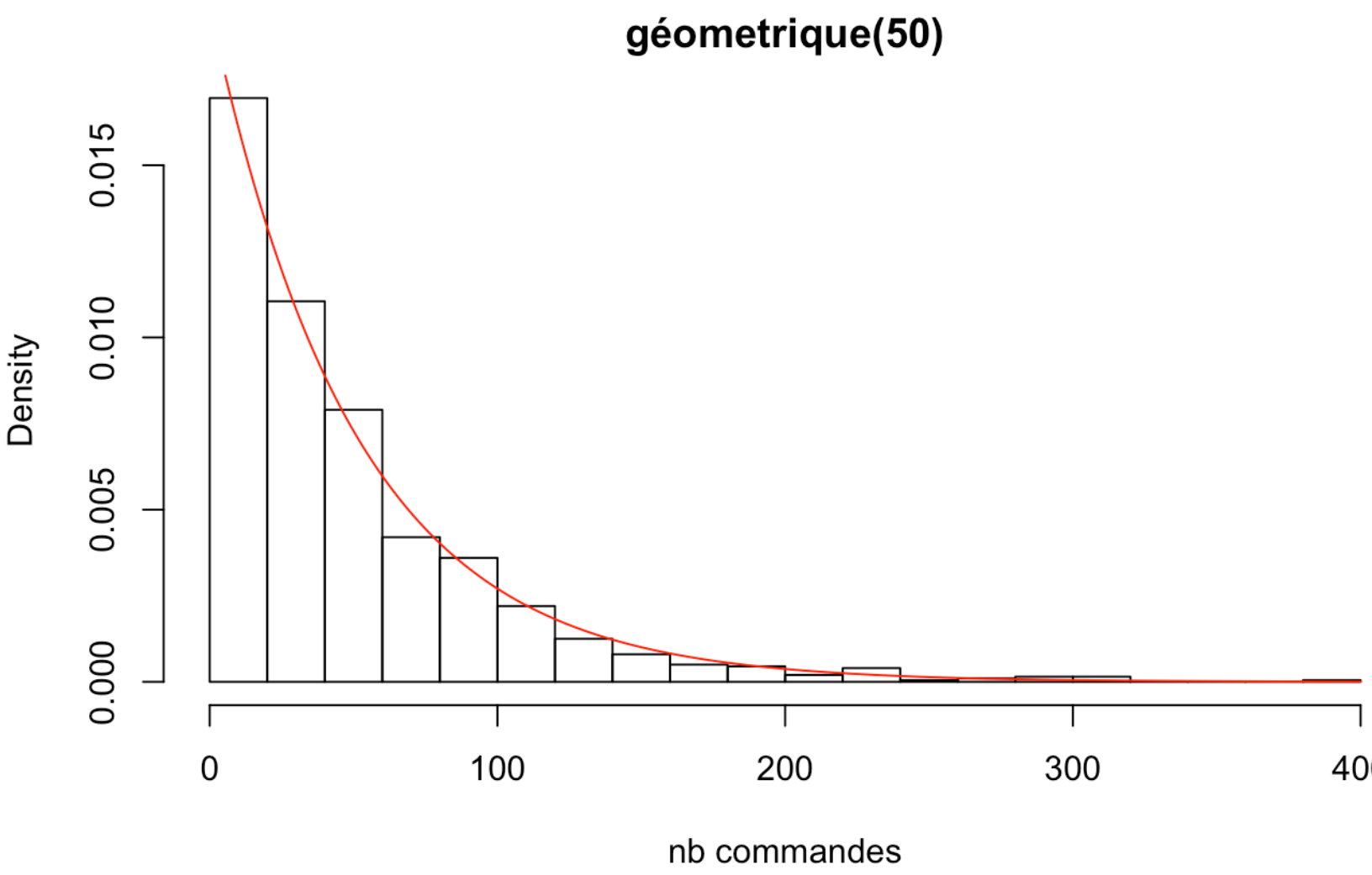
$$L(n_r, n_b, n_v, d = i \,|I_2) = \sum_{n_r=0}^{\infty}(n_r - 200\delta_{ir} - 100)\frac{1}{51}\left(\frac{50}{51}\right)^{n_r} +$$

$$\sum_{n_b=0}^{\infty}(n_b - 200\delta_{ib} - 150)\frac{1}{101}\left(\frac{100}{101}\right)^{n_b} + \sum_{n_v=0}^{\infty}(n_v - 200\delta_{iv} - 10)\frac{1}{11}\left(\frac{10}{11}\right)^{n_v}$$

Hide

Hide

```
geo=function(p,n){(1-p)*p^n}
#lambda=p/1-p, lambda est la moyenne
#p=lambda/(1+lambda) , p est la proba de survie
geom=function(lambda,n){geo(lambda/(1+lambda),n)}
# loi exponentielle discrète
rgeom=function(size,lambda,nmax=10000){
  n=0:nmax
  sample(x=n,size=size,replace = T,prob = geom(lambda,n))}
hist(rgeom(1000,50),freq=F,nc=25,xlab='nb commandes', main="géometrique(50)")
lines(1:400,geom(50,1:400),typ='l', col='red')
```



Expected utilities for the second state of knowledge

Hide

Hide

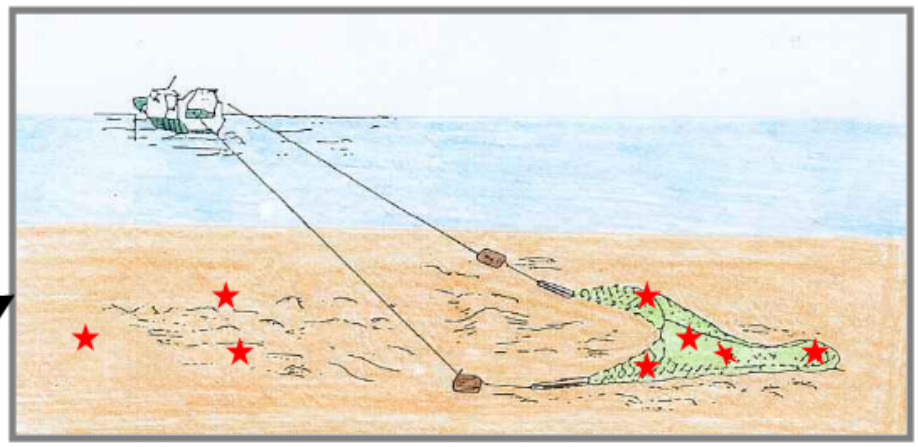
```
####Info 2
lambda=c(50,100,10)
stock=c(100,150,50)
U=c(0,0,0)
nrepet=10000
s=matrix(stock,nr=nrepet,nc=3,byrow = T)
n=matrix(0,ncol=3,nr=nrepet)
#Aleas
for (i in 1:3){
n[,i]=rgeom(nrepet,lambda = lambda[i])
}
#Decisions
for (i in 1:3){
ntest=n
ntest[,i]=ntest[,i]-200
U[i]=sum(apply(rampe(ntest,s),2,mean))
}
print(U)
```

```
[1] 22.6284  9.7273 29.2749
```

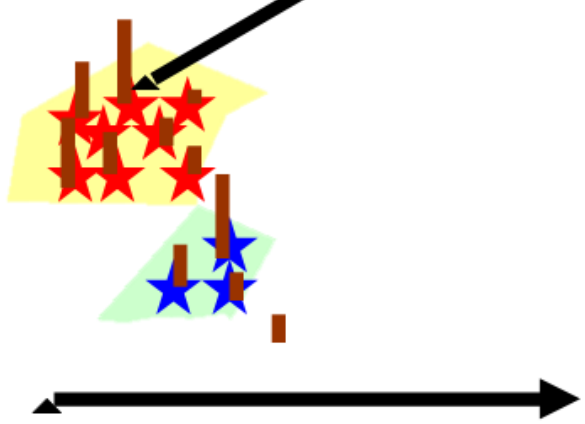
Third state of knowledge : Law of leaks

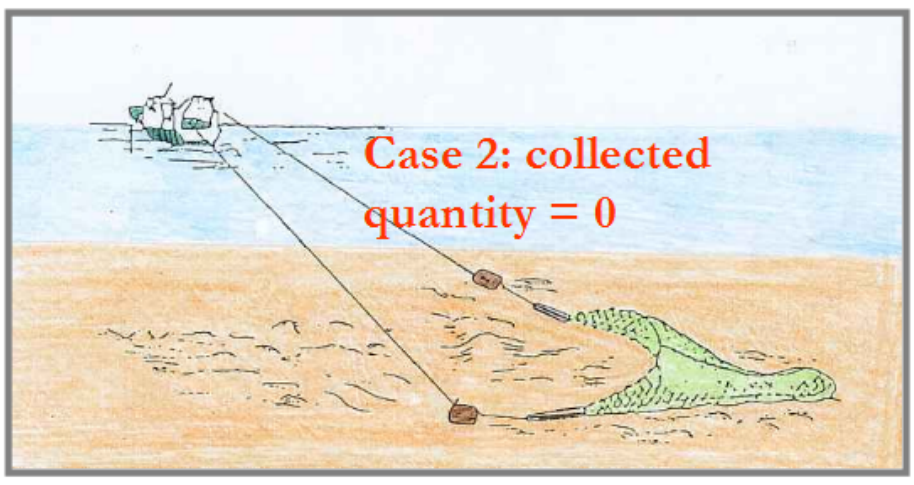
My favorite model for zero inflated data = A compound Poisson

An ecological understanding of the law of leaks



Case 1: collected quantity
=
sum of the quantities in each « clump »





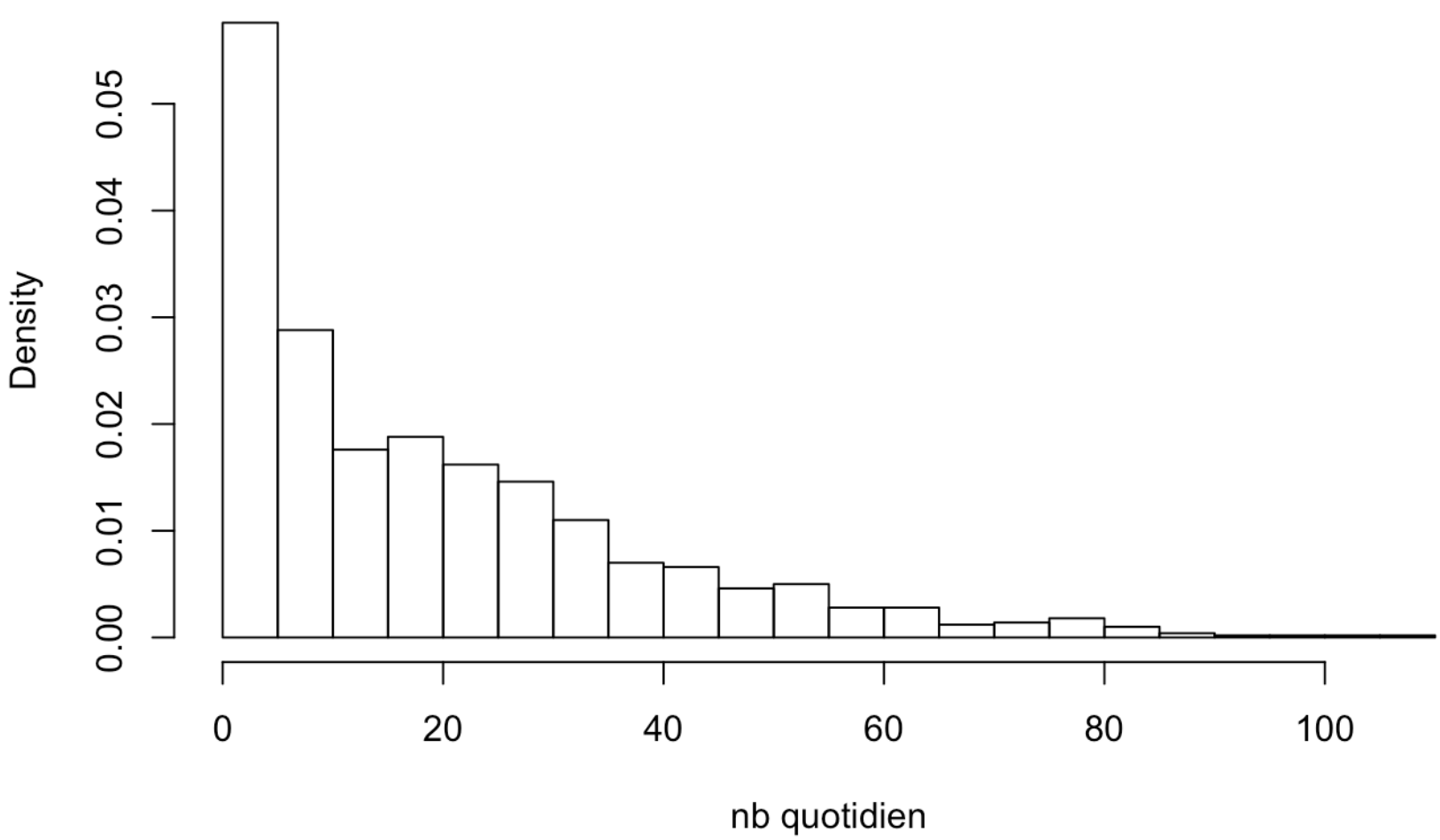
= a « clump» with an unknown random quantity

Third state of knowledge (cont'd)

Hide

Hide

```
####Info 3
lambdatot=c(50,100,10)
lambdataille=c(75,10,20)
nbComMean=lambdataille/lambdatot
stock=c(100,150,50)
U=c(0,0,0)
nrepet=1000
s=matrix(stock,nr=nrepet,nc=3,byrow = T)
n=matrix(0,ncol=3,nr=nrepet)
#Aleas
for (i in 1:3){
  n[,i]=rpois(n = nrepet,lambda = nbComMean[i])
  for (k in 1:nrepet){
    n[k,i]=sum(rgeom(n[k,i],lambda = lambdatot[i]))
  }
}
hist(n[,3],freq=F,nc=25,xlab='nb quotidien', main="fuite(2,50)")
```



Hide

Hide

```
#Decisions
for (i in 1:3){
ntest=n
ntest[,i]=ntest[,i]-200
U[i]=sum(apply(rampe(ntest,s),2,mean))
}
print(U)
```

```
[1]  5.204 23.266 23.990
```

Fourth state of knowledge

Play it again…

Hide

Hide

```
#Info 4
stock[3]=stock[3]-40
s=matrix(stock,nr=nrepet,nc=3,byrow = T)
for (i in 1:3){
ntest=n
ntest[,i]=ntest[,i]-200
U[i]=sum(apply(rampe(ntest,s),2,mean))
}
print(U)
```

```
[1] 16.196 34.258 23.990
```

To sum it up: Cost function

Quelques équations…

Fonction rampe

$$r(n,s)=\begin{cases} n-s & si \quad n>s \\ 0 & si \quad n\leq s \end{cases}$$

Perte = fonction (action,état inconnu, information)

$$L(n_r,n_b,n_v,d="rouge")=r(n_r,s+200)+r(n_b,s)+r(n_v,s)$$

Perte attendue = fonction (action, information)

$$U(d="rouge"|I)=$$

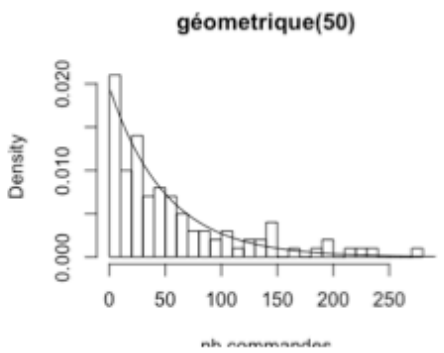
$$\sum_{n_r}\sum_{n_b}\sum_{n_v}L(n_r,n_b,n_v,d="rouge")p(n_r|I)p(n_b|I)p(n_v|I)$$

To sum it up: Conditional reasoning with probability models

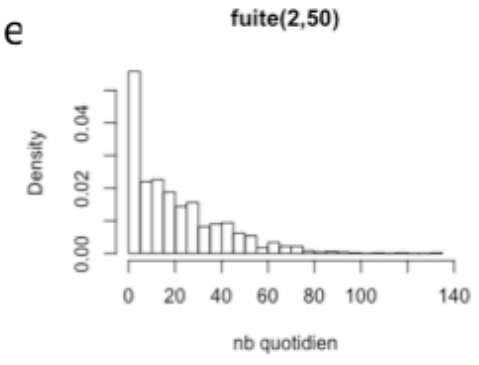
Constructions probabilistes

- Etat info 1 : $p(n_r | I_1) = p(n_b | I_1) = p(n_v | I_1)$

- Etat info 2 (Commande moyenne):
  Loi *géométrique* de moyenne fixée



- Etat info 3 (⋯+Taille moyenne de chaque commande):
  Nombre *Poissonnien* de commandes quotidiennes géométriques



Take-home message 1

- Dealing with a small sample does not mean solving a simple problem: even in that case, statistical science is intended to help us make better judgements and recommandations. Models are necessary tools to help decision making under uncertainty.

What are we relying on?

```
* specific features of knowledge (symmetry,etc.),

* principles of entropy,

* phenomenologically based interpretation,

* past experience & conveniency.
```

Part 2:
An unreliable statistical analysis and disaster.

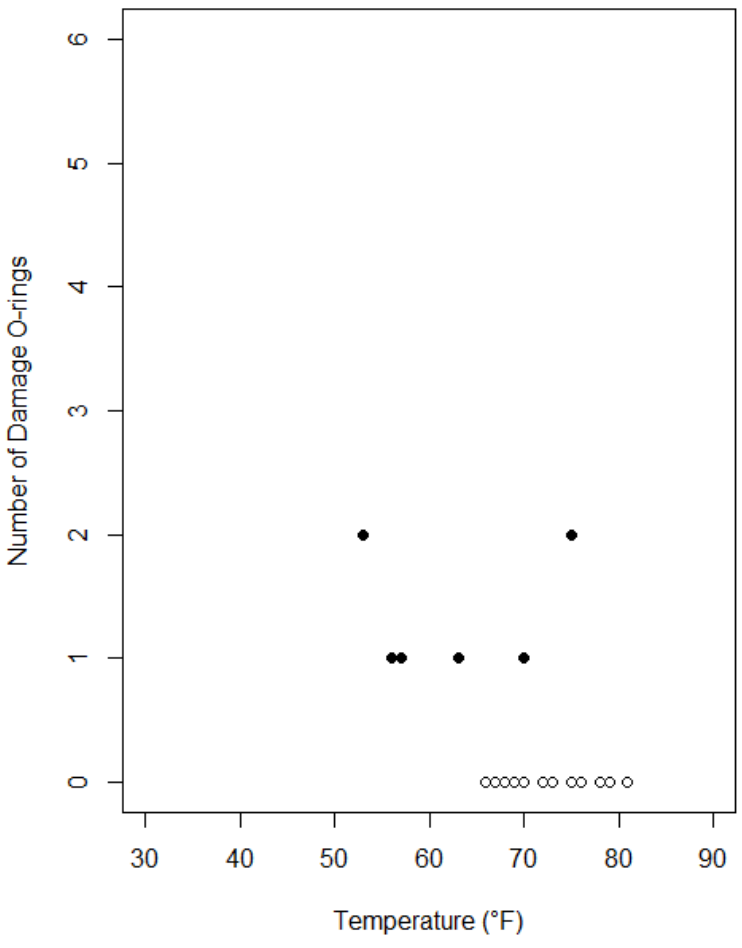
Flahback on Challenger story

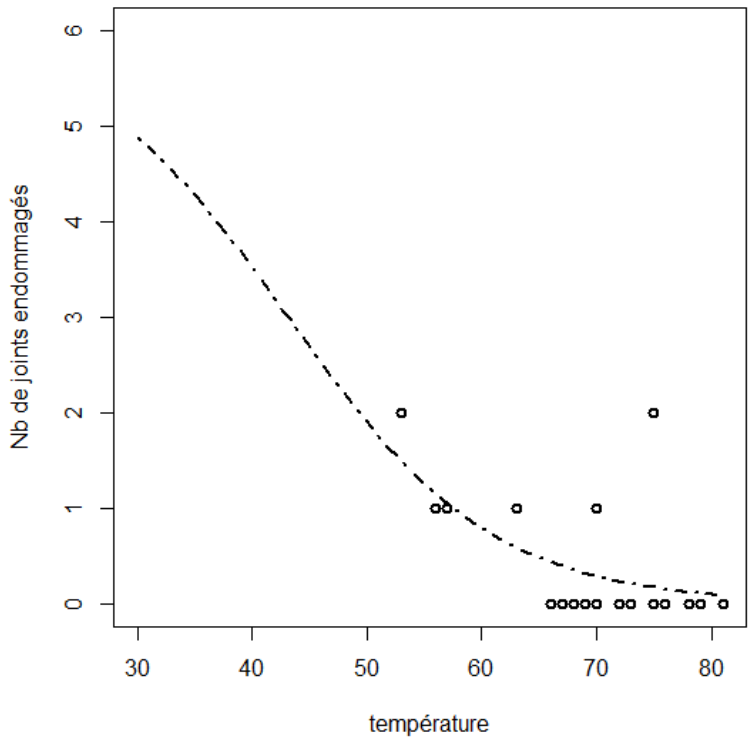
On the morning of January 28, 1986 the estimated temperature of the primary O-rings on the Challenger solid rocket motors was 31◦F (–0.6◦C). This was 22 ◦F (12.2 ◦C) lower than the minimum temperature recorded in all previous shuttle launches [◦C = (◦F – 32)5/9]. The Presidential Commission on the Space Shuttle Challenger Accident (1986) found that “a careful analysis of the flight history would have revealed the correlation of O-ring damage in low temperature.”

Damaged O-rings during previous launchings

temperature	Endommages	Pas de Dommage
53	2	4
56	1	5
57	1	5
63	1	5
66	0	6
67	0	6
67	0	6
67	0	6
68	0	6
69	0	6
70	0	6
70	0	6
70	1	5
70	1	5
72	0	6
73	0	6
75	0	6
75	2	4
76	0	6
76	0	6
78	0	6
79	0	6
81	0	6
Total	9	129

Data analysis of damaged O-rings





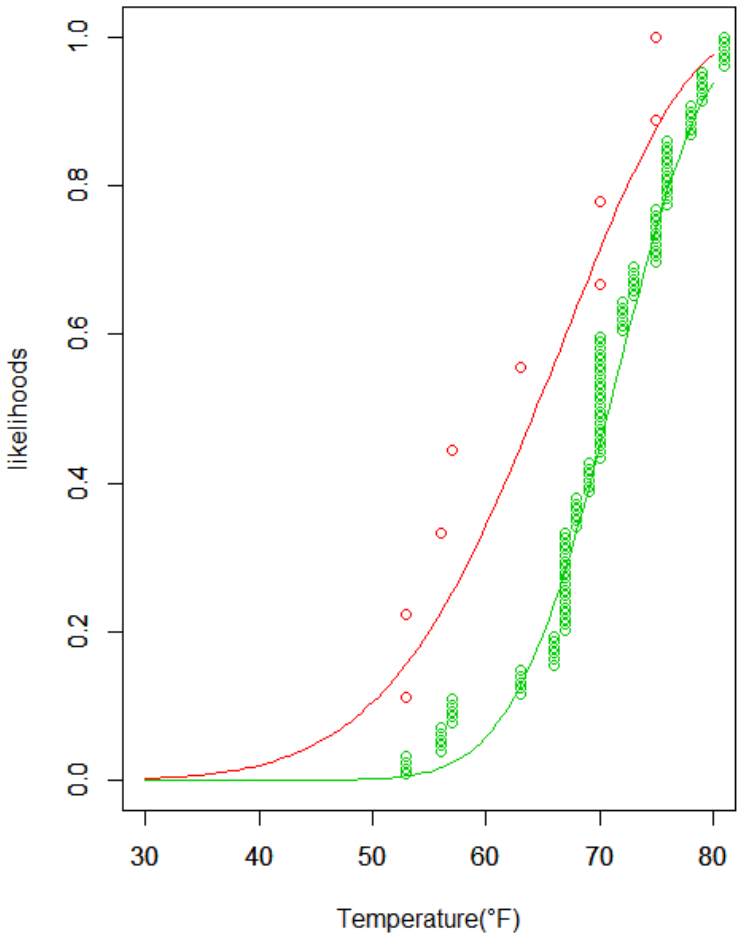
Denote X the variable temperature, $Y = 1$ if the O-ring is damaged else $Y = 0$

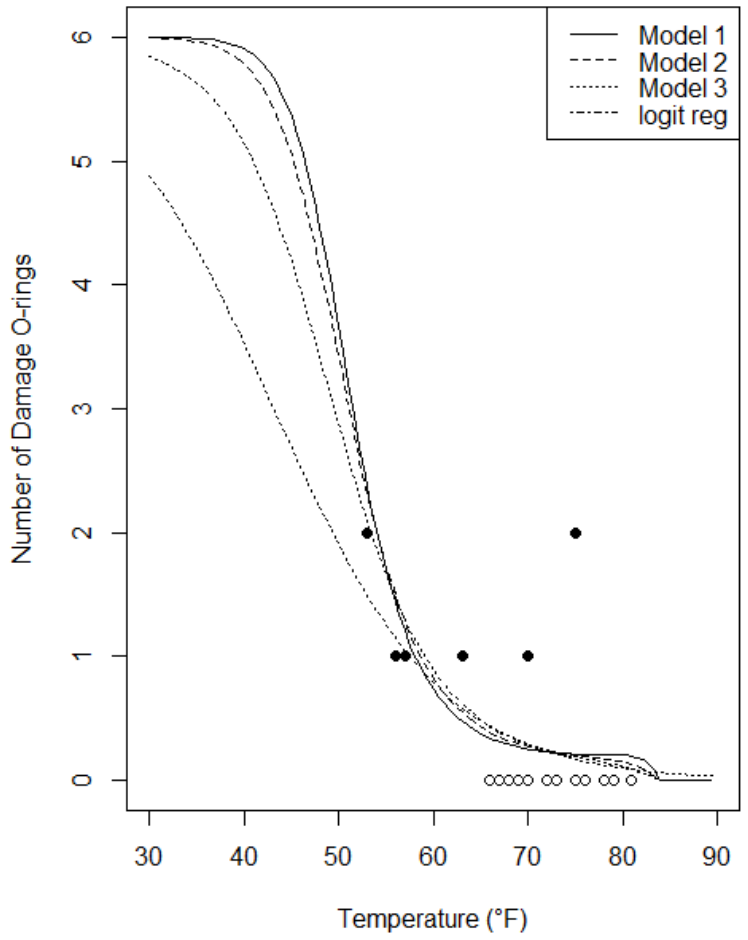
Statistical business as usual : logistic regression!

$$Y \sim Bernoulli(p)$$

$$p = \frac{exp(aX + b)}{1 + exp(aX + b)}$$

Reanalysis: in search for the joint $[X, Y]$





- try likelihoods: $[X|y = 1]$ (Weibull); $[X|y = 0]$ (Log-Weibull)
- pick a prior: $[Y = 1] = \frac{9}{9+129}$; $[Y = 0] = \frac{129}{9+129}$
- get the posterior: $[Y = 1|X] = \frac{[X|Y=1][Y=1]}{[X|Y=1][Y=1]+[X|Y=0][Y=0]}$

Take-home message 2

- Cast more than one quick eye on the figures: *The engineers at Morton Thiokol transmitted a facsimile to NASA stating that "temperature data [are] not conclusive on predicting primary O-ring blowby."*
- Be mindful of your systematic reflexes: Forcing $[Y|x]$ to be a logistic model arbitrarily removes uncertainty: the symmetry of the logistic function is such that high temperatures specify the behaviour for the (missing) low temperatures. Mind your statistical course!
- Proposing various constructions $[X|y]$, evaluating$[Y]$ and then calculating[Y|X] makes it easier to express uncertainty. Perform sensitivity analysis.

Conclusions

The contingency of small numbers

When facing a small sample, there is actually no valid narrative that can accurately explain the phenomenon one could get from the figures derived from the data set.

It is merely a matter of statistics when studying a small sample set: by nature, they are highly variable. But rather than attribute its *perceived salient* traits to the random nature of small sample sets, we cannot avoid setting up to the task of generating a story that can explain what we see.

Models convey information

- modeling = the art of cutting,
- model as a prior expertise brings information into the analysis.

Being honest as a statistician ?

- Make explicit the model assumptions : (*hypo–thesis*)
- Point out the sources of uncertainty
- Try to quantify your credibility in each proposal
- Beware of your model : make sensitivity analysis on your *priors*, including model structure
- Bayesian approach is not only structurally optimal, but also practically advantageous.