“SCIENCE: IMAGe IN AcTION”

7th International Workshop Data Analysis in Astronomy

LIVIO SCARSI AND VITO DI GESÙ

Editore Majorana
Foundation and Centre for Scientific Culture
FMCSC
Erice, Italy, 15-22 April 2011.

Statistical Information & *Pré-Posterior* Analysis.

Carlos Alberto de Bragança Pereira

Dept. Estatística & Núcleo de Bioinformática – IME/USP

Abstract

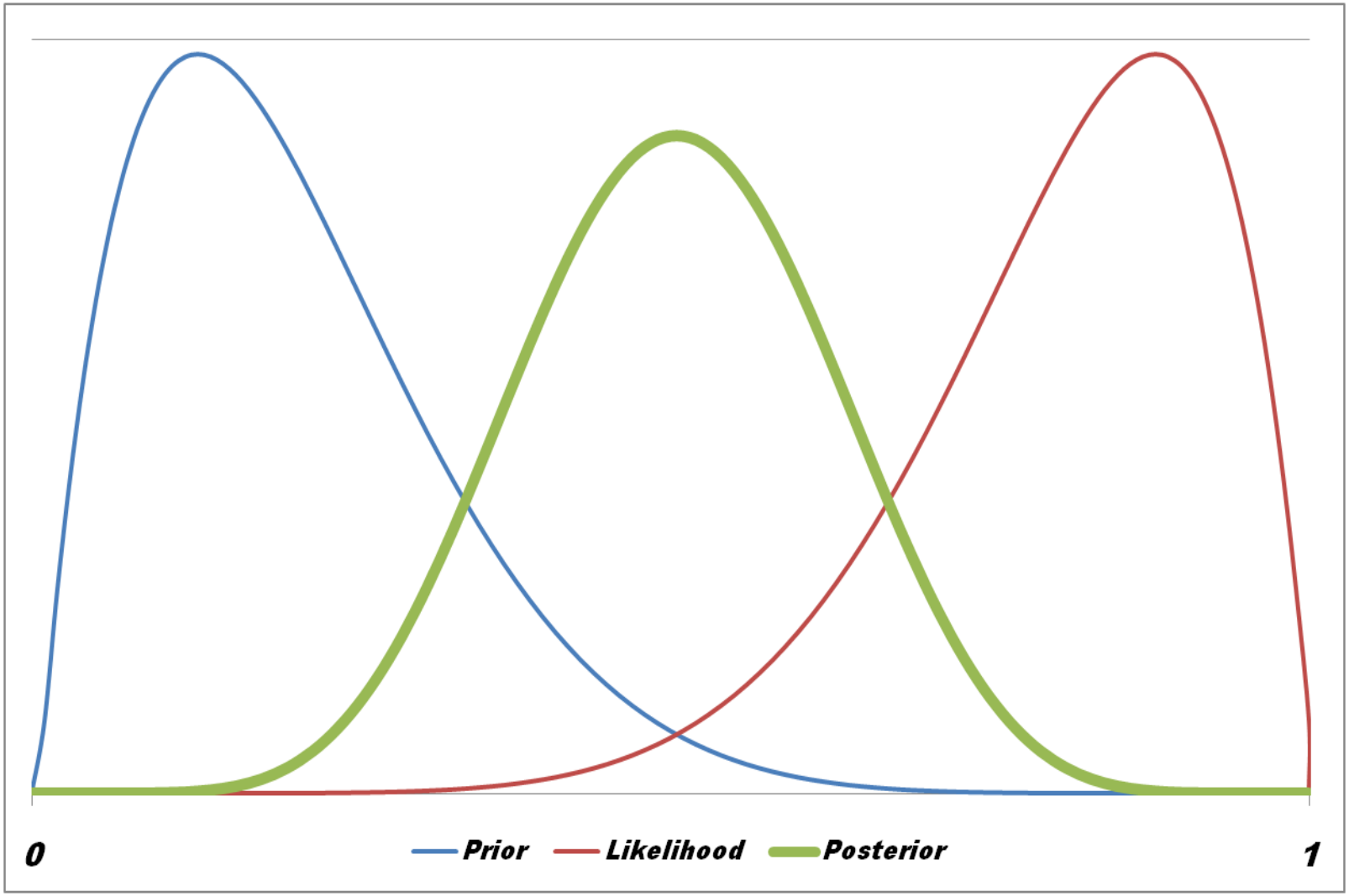
- **We explore the concept of information in statistics: information about unknown quantities of interest, the parameters.**
- **An intuitive idea of what should be information in statistics is discussed.**
- **Some tentatively measures of information are also illustrated.**
- **Operationally, information could be in the observed data or, as expected information, in a future experiment.**
- **Intuition is "proved" to be right for some interesting and common examples.**

Chalanges

- **Which experiment you would choose to provide inferences about π, an unknown proportion?**

  $Y|\pi \sim Ber(\pi)$ or $X|\pi \sim Ber(\pi/2)$?

- **You must guess the unknown number of white balls in an urn. Would you select two balls with or without replacement?**


0
Prior
Likelihood
Posterior
1

The main object of the work is a Parameter or State of Nature for which its *value*, θ, is invisible at the moment the work of the statistician starts.

A probability distribution over Θ is considered to describe our level of *uncertainty* about the value of θ.

- $P(\theta)$ -

Collecting additional *information* about θ with the objective of decrease the uncertainty, is part of our work.

Looking for a definition, I start to consider the concept presented by Basu (1975). Although not operational it seemed to be the one that best describes what most people think about INFORMATION.

D Basu (1975), Statistical information and likelihood, *Sankhyã* A 37:1-71. Lecture notes in statistics 45-SV

Information is what it does:

It changes your opinion!

The subjective aspect of the above concept is intrinsic with the inclusion of the person that is looking for additional Information. There are situations when a set of observations do not change a person opinion but drastically change the other person ones. The second person has, maybe, a different cultural information.

To operationally use the concept we must find answers to The following questions:

i. **Information: about what?**
ii. **Information: where can be found?**
iii. **Information: how much is available?**
iv. **Information: how does it can be extracted?**

Information is about the invisible value of θ.
Information is described by the actual distribution of θ.
This description is based on probabilistic evaluations
Additional information can be cultural or experimental.
The greater the involvement in the research, that looks for θ, the better is the representation of the uncertainty.
In fact, cultural information may produce a better extraction of information from observations θ.

Cultural information is *allocated* in our minds!

Experimental information is allocated in the results of an experiment. For instance, *X, Y, Z.* The process of incorporating experimental information depends on a training program, that is different from the cultural information.

This paper is about experimental working for updating distributions and consequently INFORMATION.

Let X being observed

and resulting x.

Hence calibrate from

$P(\theta)$ to $P(\theta \mid x)$.

Record this calibration

of uncertainty (probability)

is obtained by Bayes operation:

$$P(\theta \mid x) \propto L(\theta \mid x) \times P(\theta)$$

1. Information: about what?

Response: about θ.

2. Information: where?

R.: in $L(\theta \mid x)$.

3. Information: how much?

R: Distance { $P(\theta)$; $P(\theta \mid x)$ }.

4. Information: how to be extracted?

R: Using Bayes operation –

$$P(\theta \mid x) \propto L(\theta \mid x) \times P(\theta)$$

Aitchison´s compositional distance.

Let two X and Y be two k component vectors with fixed totals (it can be different totals). Our interest is only the diference in composition.

D(X;Y) = Standard deviation of L_i

$L_i = \ln(x_i)-\ln(y_i)$ & $M = (L_1+...+L_k)/k$

$D(X;Y)=\{(L_1-M)^2+...+(L_k-M)^2\}^{.5}$

Example: Consider *4* marbles, *2* transparent and *2* green.

I choose *3* and drop in a box. One must guess the number of transparent in the box.

You can, to start, take one randomly from the box to *gain* additional information.

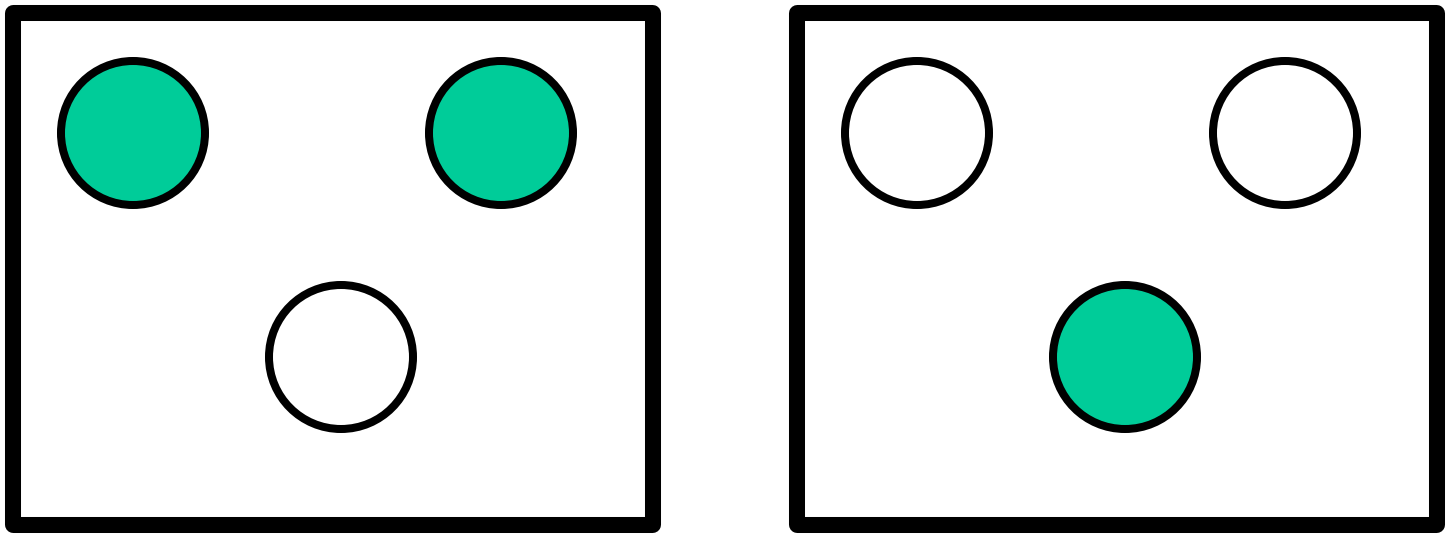
To still *get* more information, you may take a 2^{nd} Marble from the box.

This second drawing could be either with or without replacement of the 1^{st}.

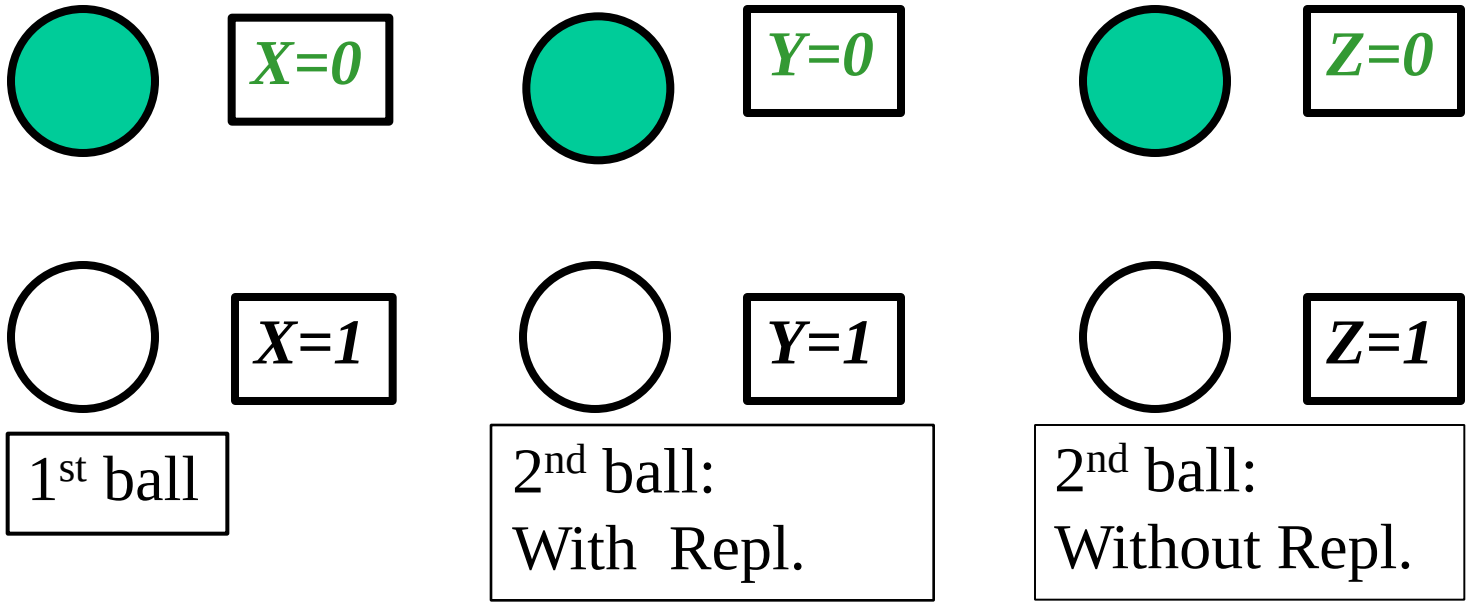
Possible States of Nature

$\theta = 1$

$\theta = 2$



Possible Samples



We must choose among

X, (X,Y)* or *(X,Z)

to be observed.

State of Nature	Prior P(θ)	Lik L(θ|x)		Post P(θ|x)	
		L(θ|0)	L(θ|1)	P(θ|0)	P(θ|1)
$\theta = 1$	0,5	2/3	1/3	2/3	1/3
$\theta = 2$	0,5	1/3	2/3	1/3	2/3

Likelihood L(θ|x,y)			
L(θ|0,0)	L(θ|0,1)	L(θ|1,0)	L(θ|1,1)
4/9	2/9	2/9	1/9
1/9	2/9	2/9	4/9

Likelihood L(θ|x,z)			
L(θ|0,0)	L(θ|0,1)	L(θ|1,0)	L(θ|1,1)
1/3	1/3	1/3	0
0	1/3	1/3	1/3

Posterior P(θ|x,y)			
P(θ|0,0)	P(θ|0,1)	P(θ|1,0)	P(θ|1,1)
4/5	1/2	1/2	1/5
1/5	1/2	1/2	4/5

Posterior P(θ|x,z)			
P(θ|0,0)	P(θ|0,1)	P(θ|1,0)	P(θ|1,1)
1	1/2	1/2	0
0	1/2	1/2	1

Probability

Predictive	*Y=0*	*Y=1*	*Z=0*	*Z=1*	*X*
X=0	**5/18**	**4/18**	**1/6**	**1/3**	**1/2**
X=1	**4/18**	**5/18**	**1/3**	**1/6**	**1/2**

Probability

Information	*X*	*(X,Y)*	*(X,Z)*
Yes	**1**	**10/18**	**6/18**
No	**0**	**8/18**	**12/18**

Ganho

Absoluto	*Y=0*	*Y=1*	*Z=0*	*Z=1*	*X*
X=0	**9/30**	**0**	**15/30**	**0**	**5/30**
X=1	**0**	**9/30**	**0**	**15/30**	**5/30**

Relativo	*Y=0*	*Y=1*	*Z=0*	*Z=1*	*X*
X=0	**60%**	**0%**	**100%**	**0%**	**33%**
X=1	**0%**	**60%**	**0%**	**100%**	**33%**

Conservative person chooses *X*, little gain but guarantied.

Choosing *(X,Y)* one may gain more but with the risk of no gain.

With *(X,Z)* one may gain the maximum, but with a high risk of no gain.

DeGroot´s Information

Here is an important paper of Prof. Morris DeGroot:

M DeGroot (1962), Uncertainty, Information, and Sequential Experiments, *Ann. Math. Statistics* 33:404-19.

Recalling the example, difficulty was to choose the best among the 3 competing experiments,

X, (X,Y) e (X,Z).

We had no reasonable criterion!

DeGroot considered an *Uncertainty*, U, defined over the set of *all* probability (density) functions taking values on the real line.

U is like an index to measure the uncertainty imbedded in the probability function, P, of θ. A solely restriction for P is as follows:

Let x be an experiment related to θ, the parameter of interest.
P & P_x are probability funcions; *prior* & *posterior*.
If U is the uncertainty function & E the expectation then:

$$U(P) \geq E\{U(P_x)\}.$$

One expects the uncertainty decreases from prior to posterior.

The following is the main result of the paper:

$$I\{x,P,U\} = U(P) - E\{U(P_x)\} \geq 0 \Leftrightarrow U \text{ is concave.}$$

The I operator is the available information about θ "contained" in X when P is the prior & U the uncertainty

We choose the experiment with the largest value of I.
The variance of θ is the best known uncertainty function.

$$I\{x,P,V\} = V(\theta) - E\{V(\theta|x)\} = V(E\{\theta|x\}) \geq 0$$

For the example one gets

$$I\{X,P,V\} = .03 < I\{(X,Y),P,V\} = .05 < I\{(X,Z),P,V\} = .08$$

The experiment with the largest variance of the Bayes estimator is the most informative and must be the one to be chosen. Note that we are looking for maximum variance, not minimum like frequentists.

Blackwell´s Sufficiency

For Basu, Blackwell´s sufficiency is the Bayesian sufficiency.
We will discuss this in the sequel.
Blackwell´s sufficiency was created to generalize the standard concept of Fisher´s sufficiency. Both compare random variables.

Fisher´s were for ones in the same sample space.
Blackwell´s abandon such restriction.

This is the main reference:
D Blackwell (1951), Comparison of experiments,
in: ***Proc. of the 2nd Berkeley Symposium*, 93-102**.

Basu was looking to attend the following property:
It suffices to observe X in the place of Y for inferences about θ if, for every possible y of Y, there exists a sample point x of X such that $P(\theta|Y=y) = P(\theta|X=x)$. I. e.,

$$\forall y \in Y, \exists x \in X \ \ st \ \ P(\theta|Y=y) = P(\theta|X=x).$$

If this property holds for all priors, this is in fact a Likelihood property, maybe equivalent to Blackwell sufficiency.

The posterior is obtained from the likelihood, obviously.

$$L(\theta|y) \propto L(\theta|x).$$

It seems that Blackwell Sufficiency goes together with the Likelihood principle.
One expects: its use do not violate the principle.

We do not make any distinction among experiments and random variables

Let X **&** Y **with respective sample spaces be** $\mathcal{X}$ **&** $\mathcal{Y}$ **.**
A transition function, F**, from** $\mathcal{X}$ **to** $\mathcal{Y}$ **is a family**

$$F = \{f_x(.);\ x \in \mathcal{X}\}$$

of probability (density) functions $f_x(y)$
over $\mathcal{Y}$ **indexed by** $x \in \mathcal{X}$ **.**

For example the family of Hipergeometric probability $H(y;x,n,N)$ **is a transition function from** $\{0,1,...,N\}$ **to** $\{0,1,...n\}$**,** $n < N$ **being positive integers.**

Definition:
Let X & Y two experiments, as above, having probability (densities) functions

$$g(x|\theta) \quad \& \quad h(y|\theta).$$

X is sufficient for Y, respect to θ, in the sense of Blackwell, if there is a transition function such that

$$h(y|\theta) = \Sigma_x f_x(y) g(x|\theta).$$

To understand that the Sufficiency of Blackwell is really a Bayesian one lets take an example.

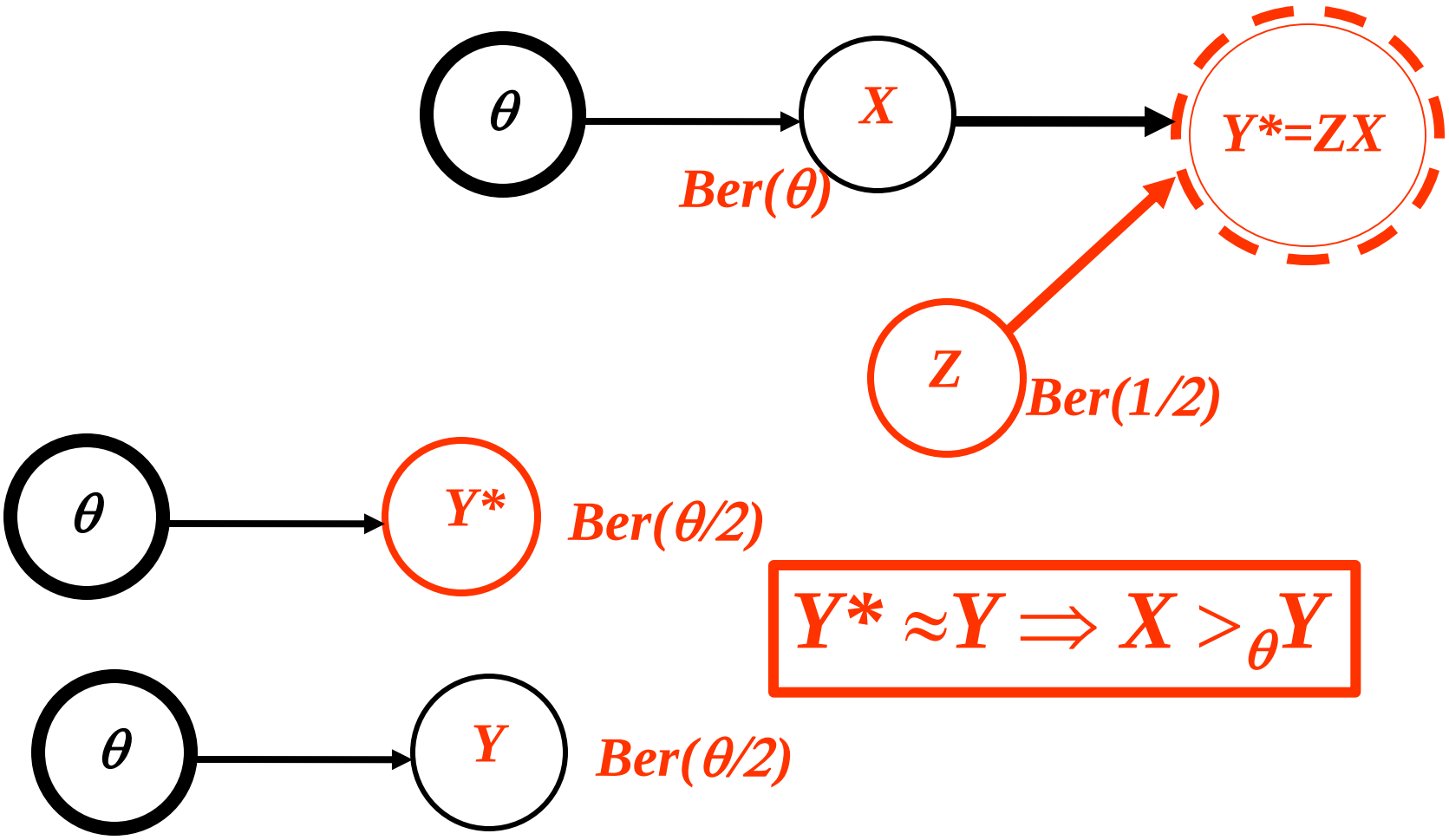
Example: A company claims that the failure rate of their product is half of the rate θ of another
To estimate θ one could take samples from either one
That is, one must choose between

$$X|\theta \sim Ber(\theta) \;\&\; Y|\theta \sim Ber(\theta/2).$$

$$X >_\theta Y,\ Y >_\theta X \text{ or } X \approx_\theta Y \ ?$$

D Basu & CA de B Pereira (1990), Blackwell Sufficiency and Bernoulli Experiments, *BJPS* 4:137-45.

To understand that $X >_\theta Y$ consider the influence diagram.



$$Y^* \approx Y \Rightarrow X >_\theta Y$$

Teorema:

Let $X|\theta \sim Ber[f(\theta)]$ & $Y|\theta \sim Ber[g(\theta)]$.

X & Y are comparable in the sense of Blackwell if the family

$$\{(f(\theta), g(\theta)): \theta \in \Theta\}$$

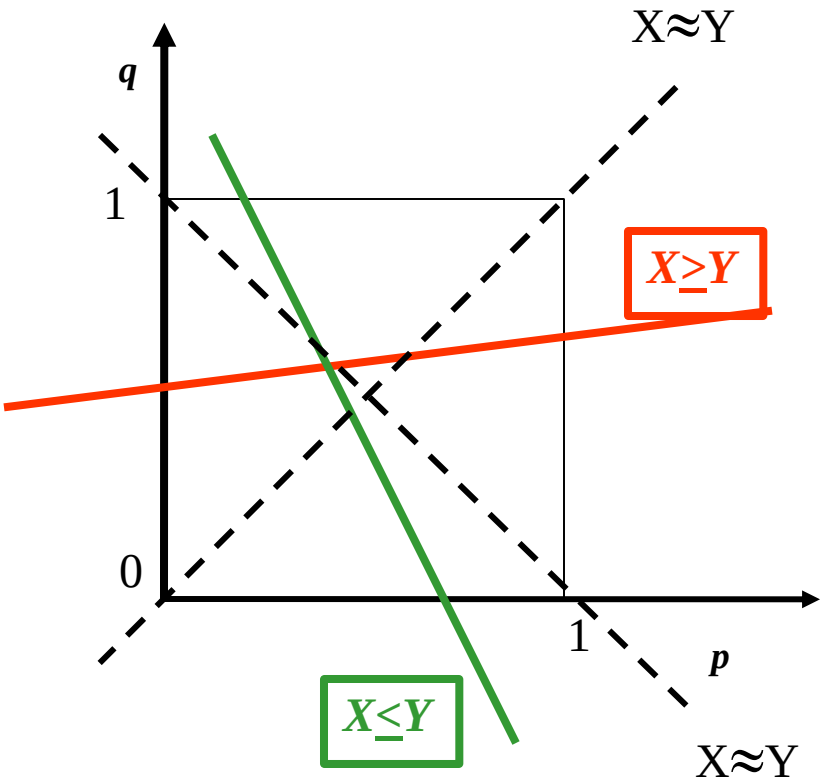
is contained in a line that cuts two opposed sides of the unity square $[0,1]^2$:

$$X >_\theta Y \quad (Y >_\theta X)$$

for the vertical (horizontal) sides.

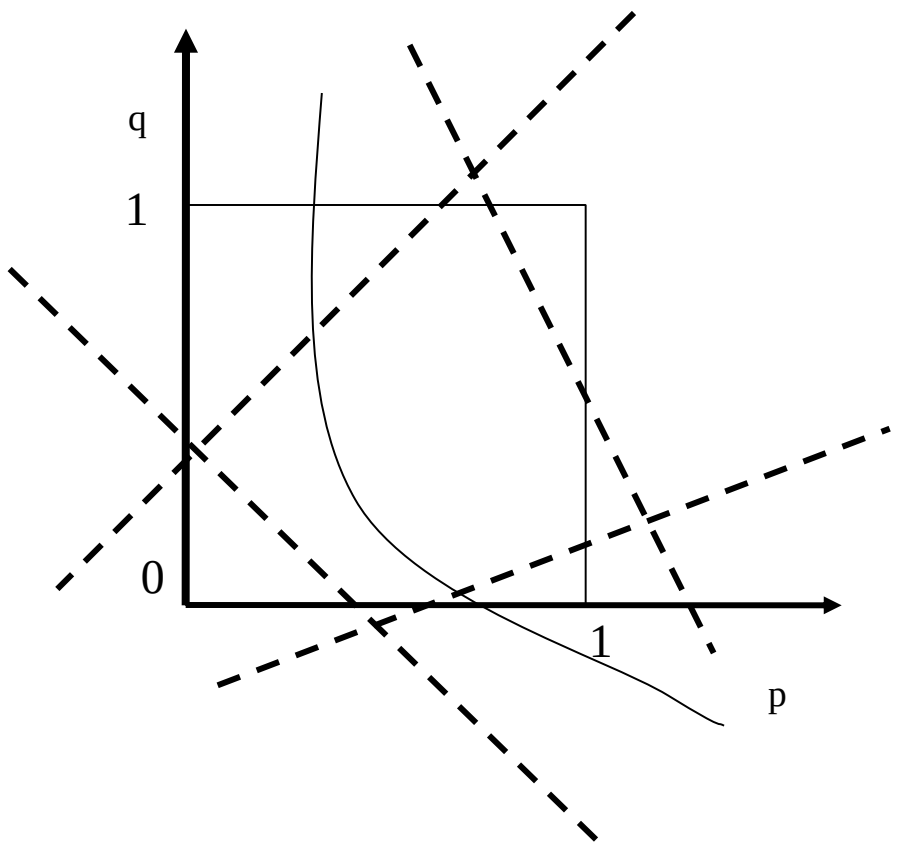
For the diagonal line $X \approx_\theta Y$.

Recalling the first example of the marbles, we can show that $Z >_\theta Y$. I. e., sample without replacement is more INFORMATIVE than sample with replacement.



Comparable

Not comparable



For Bernoully, X e Y, the existence of transition function corresponds to the existence of a transition matrix.
That is, let us take X sufficient to Y.
In symbols

$$P(Y=1|\theta) = P(X=0|\theta)f_0(1) + P(X=1|\theta)f_1(1).$$

That is,

$$h(\theta) = (1-g(\theta))f_0(1) + g(\theta)f_1(1) = f_0(1) + g(\theta)[f_1(1) - f_0(1)]$$

$$1-h(\theta) = [1-g(\theta)][1-f_0(1)] + g(\theta)[1-f_1(1)]$$

That is: $[1- h(\theta) ; h(\theta)] = [1- h(\theta) ; h(\theta)]F$

$$F = \begin{pmatrix} 1- f_0(1) & f_0(1) \\ 1- f_1(1) & f_1(1) \end{pmatrix}$$

As a consequence of the result we have that
For $0 < \theta < 1$, comparing

$X|\theta \sim Ber[\theta]$ & $Y|\theta \sim Ber[c\theta]$.

$X >_{\theta} Y$ ($Y >_{\theta} X$) in the case of $0 < c < 1$ ($c > 1$).

Consider a box with N marbles from which θ are transparent. Take the first marble and record if it is transparent. Denote the experiment by $X|\theta \sim Ber[\theta/N]$.
To continuing sampling, our challenge is to decide replacing or not the first marble into the box.
As before, Y (Z) is relative to the experiment with (without) replacement.
The Influence Diagrams are.

$P(Y=1|X=0,\theta) = P(Y=1|X=1,\theta) = P(Y=1|\theta)=\theta/N$

$P(Z=1|X=0,\theta) = \theta/(N-1)$ & $P(Z=1|X=1,\theta) = (\theta-1)/(N-1)$

For $X=0$,

$f(1|\theta)= 0g(0|\theta) + [(N-1)/N]g(1|\theta)$

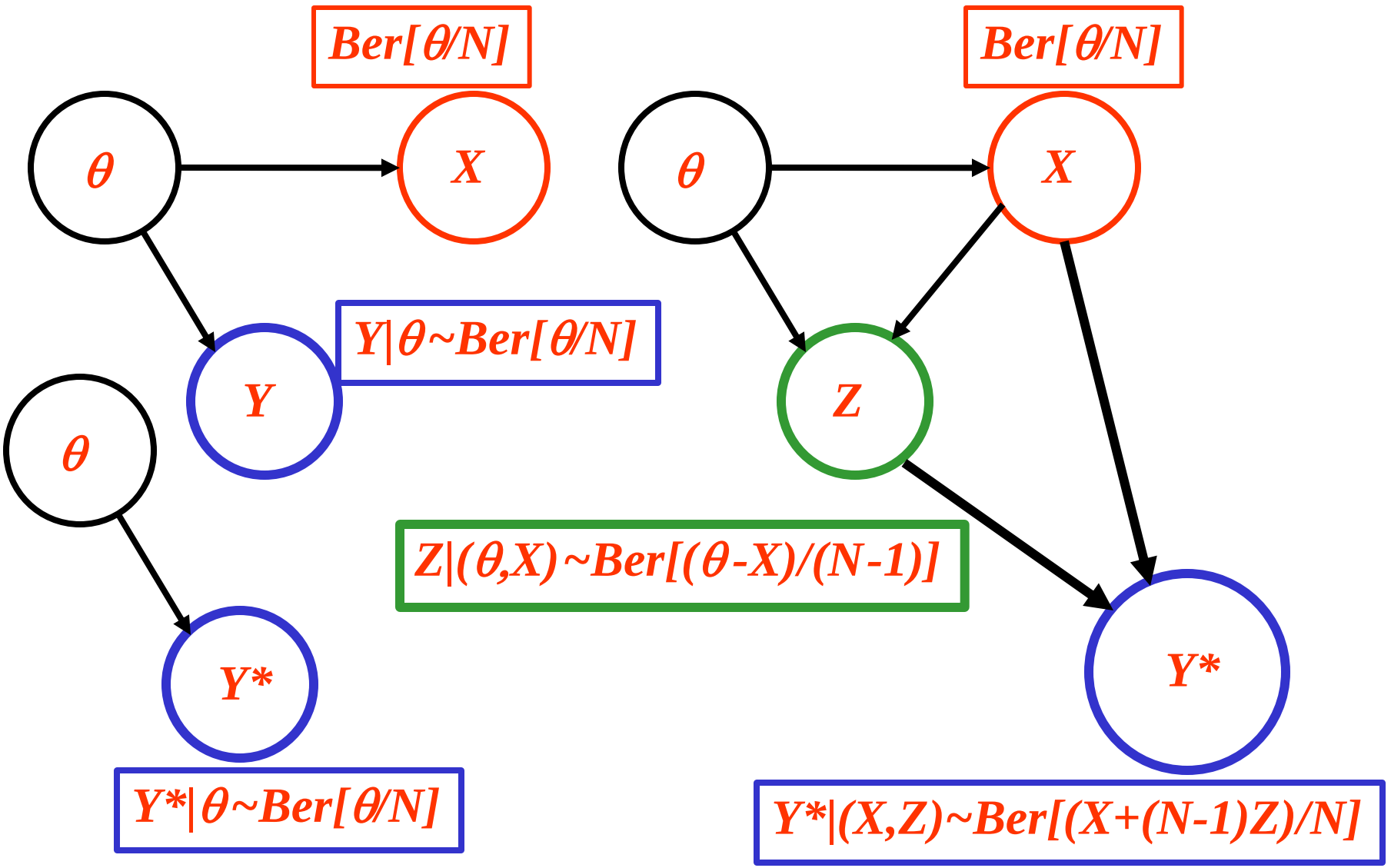
$f(0|\theta)=1-\theta/N = [1- \theta/(N-1)]+(1/N)[\theta/(N-1)]=$
$=1g(0|\theta)+(1/N)g(1|\theta)$

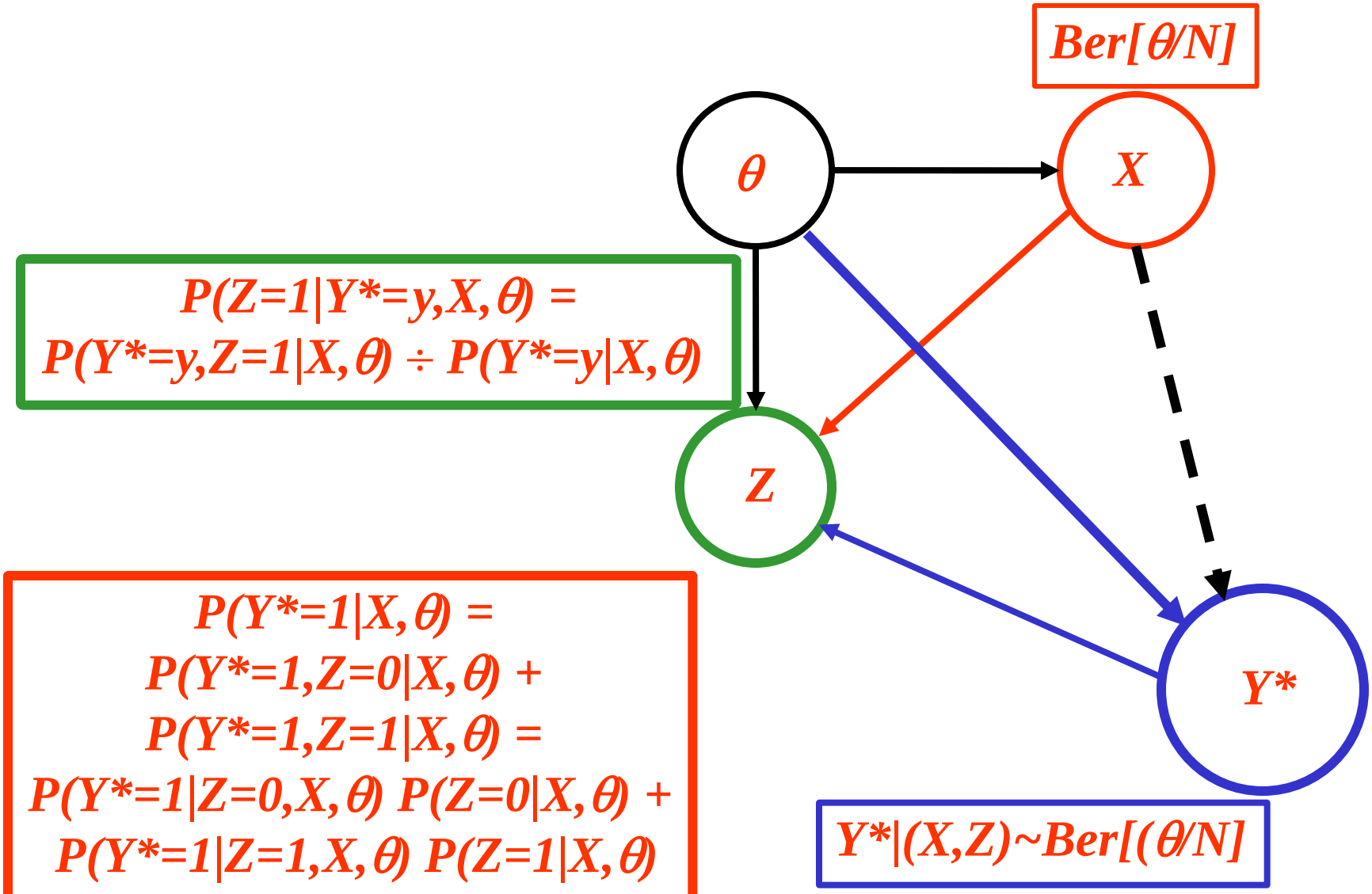
For $X=1$,

$f(0|\theta)= [(N-1)/N]g(0|\theta) + 0g(1|\theta)$

$f(1|\theta)= \theta/N = (1/N)[(N-\theta)/(N-1)]+ (\theta-1)/(N-1)=$
$= (1/N)g(0|\theta) + 1g(1|\theta)$

$$T_0 = \begin{pmatrix} 1 & 0 \\ \frac{1}{N} & \frac{N-1}{N} \end{pmatrix}; \quad T_1 = \begin{pmatrix} \frac{1}{N} & \frac{N-1}{N} \\ 1 & 0 \end{pmatrix}$$


Ber[θ/N]
θ
X
Y|θ~Ber[θ/N]
Y
θ
Y*
Y*|θ~Ber[θ/N]
Ber[θ/N]
θ
X
Z
Z|(θ,X)~Ber[(θ-X)/(N-1)]
Y*
Y*|(X,Z)~Ber[(X+(N-1)Z)/N]


Ber[θ/N]
θ
X
Z
Y*
P(Z=1|Y*=y,X,θ) =
P(Y*=y,Z=1|X,θ) ÷ P(Y*=y|X,θ)
P(Y*=1|X,θ) =
P(Y*=1,Z=0|X,θ) +
P(Y*=1,Z=1|X,θ) =
P(Y*=1|Z=0,X,θ) P(Z=0|X,θ) +
P(Y*=1|Z=1,X,θ) P(Z=1|X,θ)
Y*|(X,Z)~Ber[(θ/N]

M Skibinsky (1970), **A characterization of hypergeometric distributions,** ***JASA*** **65:926-29.**

D Basu & CA de B Pereira (1983), **A note on Blackwell sufficiency and a Skibinsky characterization of distributions,** ***Sankhyã A*** **45(1):99-104.**

Using completeness and Blackwell sufficiency we could characterize as follows:

1. **Multinomial as transition functions for Poisson**
2. **Hipergeometric for multinomials.**
3. **Dirichelet-Multinomial for Negative Binomials c**

R.1: x & y vetors of positive integers such that $1'x = N$ & $1'y = n$. On the other side, $x|\theta \sim M_k(N,\theta)$ e $y|\theta \sim M_k(n,\theta)$.

$$f(y|\theta) = \sum_x h_x(y)g(x|\theta) \Leftrightarrow h_x(y) \sim H(N,n;x).$$

R.2: r a ***k-vector*** such that $0 < r < 1$ e $1'r=1$.

y a k-vector of independent rv such that $y_i|\theta \sim Po(\theta r_i)$. If $x|\theta \sim Po(\theta)$ then,

$$f(y|\theta) = \sum_x h_x(y)g(x|\theta) \Leftrightarrow h_x(y) \text{ pf of } M_k(1'x\ ;r).$$

R.3: r a ***k-vector*** of ***positive*** integers.

y a k-vecor of independent rv's such that $y_i|\theta \sim NB(\theta; r_i)$. If $x|\theta \sim NB(\theta; 1'r)$ then,

$$f(y|\theta) = \sum_x h_x(y)g(x|\theta) \Leftrightarrow h_x(y) \text{ pf of } DM_k(1'x\ ;r).$$

classes	E	E´	total
F	θ	$f-\theta$	f
F´	$e-\theta$	$1-e-f+\theta$	$1-f$
total	e	$1-e$	1

$$0 < e < f < 1-f < 1-e < 1$$

$$X|\theta \sim Ber[\theta];$$

$$X_E|\theta \sim Ber[\theta/e]; \qquad X_{E'}|\theta \sim Ber[(f-\theta)/(1-e)]$$

$$X_F|\theta \sim Ber[\theta/f]; \qquad X_{F'}|\theta \sim Ber[(e-\theta)/(1-f)]$$

X_E *is sufficient for all experiments.*

$$X_E >_\theta X_F >_\theta X \qquad X_E >_\theta X_F >_\theta X_{E'} \qquad X_E >_\theta X_{F'} >_\theta X_{E'}$$

The other pairs are not comparable:

$$(X_F ; X_{F'}) \qquad (X ; X_{E'}) \qquad (X ; X_{F'})$$

Using completeness and Blackwell sufficiency we could characterize as follows

1. **Multinomial as transition functions for Poisson**

2. **Hipergeometric for multinomials.**

3. **Dirichelet-Multinomial for Negative Binomials**

We have now the following examples of Blackwell sufficiency:

1. If $X \sim Ber(\pi)$ & $Y \sim Ber(q\pi+(1-q)p)$, then X BS for Y.
2. If $X \sim Bin(p)$ & $Y \sim Bin(q\pi+(1-q)p)$, then X BS for Y.
3. Sampling without replacement is BS for with replacement
4. $X \sim N(\mu;s)$ and $Y(\mu;s')$ with $s < s'$: X is BS for Y.
5. $X \sim Po(\lambda)$ and $Y \sim Po(p\lambda)$ and p is in [0;1]: then X BS for Y.
6. $X \sim Ex(\lambda)$ and $Y \sim Ex(p\lambda)$ and p is in [0;1]: then X BS for Y.

Blackwell Likelihood Theorem

$X \cong Y$ *If & only if*

$$P[\{x \in \Xi; L_X(\bullet \mid x) \propto L(\bullet \mid x)\} \mid \theta] =$$

$$P[\{y \in \mathrm{Y}; L_Y(\bullet \mid y) \propto L(\bullet \mid y)\} \mid \theta]$$

$\forall\, \theta \in \Theta$ & $\forall$ *Lik function* L *derived from either* X *or* Y.