-.-   .-

Causal INSIGHTS INSIDE for data mining to fight data tsunami and confounding
or
Causation and confounding as indicated by probabilistic Implication*Surprise in relative risk RR, likelihood ratio LR, I.J.Good, Kemeny vs Popper, Google, for data mining in epidemiology, evidence-based medicine, economy, investments

Version 1.59 of May 27, 2004, has 3252 lines of < 79+CrLf chars in ASCII, likely to be updated soon, written with "CMfiler 6.06f" from www; submitted to the webmaster of http://www.matheory.info aka http://www.matheory.com .
This epaper may read better (more of left margin, PgDn/Up) outside your email.
This epaper has new facilities for fast finding and browsing. Please save the last copy and use a file differencer to see only where the versions differ: download Visual Compare VC154 and run it as: VCOMP vers1 vers2 /k /i which is the best and the brightest colorful comparer for plain .txt files.
Your comments (preferably interlaced into this .txt file) are welcome.

Browsers may like to repeatedly find the following markers :
!!! !! ! ?? ? { refs }
Q: Single spaced keywords on this list indicates semantical closeness :
?(asymmetr attributable etiologic B(B(~ Bayes factor beta Bonferroni
:-) as-if boost confound Cornfield Gastwirth caution chain conjecture
:-(Brin caus1(causa causes code cofa cofa0 cofa1 CI confidence confound
contingency cont(--> conv(conv1 conv2 conv3 conviction corr(correl cov(
confirm C(corroborat counterfactual degree depend DeMorgan entrop
error example expos F(F(~ F0(factual support Kemeny Gini I.J. Good
Hajek hypothe --> impli 0/0 independ infinit oo inhibit inh0(inh1(
Kahre Kemeny key likelihood LR meaning mislead MDL MEL MML LikelyThanNot
necess suffic Occam odds(PARADOX Pearson Phi Popper princip proper
ratio relative risk RR(RR(~ r2 refut relativi rapidit regraduat remov
rule NAIVE Schield SIMPLISTIC SeLn(SeLn(OR) SeLn(RR) SIC slope Shannon
Sheps surpris symmetr Spinoza Venn 2x2 table 5x2 tendency triviality
variance regress tanh TauB UNDESIRABLE weigh evidence W(W(~ WinRatio
WR www opeRation -log(sense exaggerat Folk Google 17th
-.- separates sections .- separates (sub)sections |- tables & Venn diagrams

+Contents : each +Word allows instant finding of the section; the content
            of each section is much better than the Contents suggest

-.-

+Who might like to read this epaper :

This epaper started as notes to myself (Descartes called them Cogitationes privatae). Now it is a much improved version of my original draft tentatively titled "Data mining = fighting the data tsunami :
When & how much the evidential event y INDICATES x as a hypothesised cause,
for doctors, engineers, investors, lawyers, researchers and scientists",
who all should be interested in this stuff. This epaper is primarily targeted at British-style empiricists or BE's (sounds better than BSE :-). Continental Rationalists (CR's) a la Descartes, Leibniz and Spinoza prefer to apply deductive analytical methods to splendidly isolated and well defined problems, while BE's a la Locke, Berkeley, Hume are not afraid of using inductive inferential/experimental/observational methods even on messy tasks in biostatistics, econometry, medicine, and in military and social domains.
BE's credo is Berkeley's "Esse est percipi".
CR's credo is Descartes' "Cogito ergo sum".

-.-

+Intro :

When confronted with events, and events happen all the time, humans ask about and search for inter-event relationships, associations, influences, reasons, and causes, so that predictions, remedies and decision-making may be learned from the past experiences of such or similar events. To find a cause, an explanation, and/or a remedy is the ultimate goal, the Holy Grail of advisors, analists, attorneys, barristers, doctors, engineers, investigators, investors, lawyers, philosophers, physicians, prosecutors, researches, scientists, and in fact of all wonderful expert human beings like you and me, who use or just think the words "because", "due to", "door" (in Dutch :-), and "if-then". David Hume (1711-1776) used to say that the "causation is the cement of the Universe". Max Planck (1858-1947) quoted in { Kahre 2002, p.187 } :
"Causation is neither true nor false, it is more a heuristic principle, a guide, and in my opinion clearly the most valuable guide that we have to find the right way in the motley hotchpotch [= bunten Wirrwarr], in which scientific reserch must take place, and reach fruitful results."

One man's mechanism is another man's black box,
          wrote Patrick Suppes at Stanford, and I say:
One man's data is another woman's noise, and also:
one man's cause is another woman's effect, eg:

```
  gene ...> hormone level ...> symptom ; or if we view the notion of specific
            illness as-if real (in fact it is an abstraction), then eg:
  gene ...> illness ........> symptom .  In this causal chain
```

a researcher may see the illness as an effect caused by genes, while a physician, GP or clinician, sees it as a cause of a symptom, eg a pain in the neck to be removed or at least suppressed. Cause-effect relationships are relative wrt to the observer's frame of view, like EinStein would have loved to say.
Like an implication, causation is supposed to be transitive like eg in math
   if A > B and B > C then A > C.

```
!!! Caution: causation works in the opposite direction wrt implication. This
             is so, because ideally an effect  y  implies a cause  x, ie
      a cause  x  is necessary for an effect  y (draw a Venn diagram).
Note that an inference rule :
          IF effect ie evidence  THEN   hypothesised cause (eg an exposure)
is reflected in the  ( evidence implies hypothetical cause (eg a treatment),
while the causation goes in the opposite direction:
( exposure may cause effect or evidence ). Hence we must be careful about the
assigned meanings and about directions of arrows and notations like (a:b),
(x:y) , (y:x) , (~y:~x) , (~x:~y) , conv(x --> y) , etc.  Many cues or
predictors are symptoms caused by a health disorder, but some cues are surely
the causes of an illness, so eg :
IF (wo)man THEN "(fe)male disorder likely"  makes sense, but it would be
  foolish to think that a disorder caused a human to be a (wo)man. Although
IF (fe)male disorder THEN (wo)man, is correct, it (usually) is pointless.

-.-

+Extended abstract = Insight inside :

.-  +Key contrasting formulas :

Too many measures of statistical association were (re)invented under even more
all too suggestive names { Feinstein 2001, sect.17.6, pp.337-340 tells many }.
All of them somehow capture statistical dependence, often in form of ARR(:) or
RR(:) which both contrast P(y|x) vs P(y|~x). The key question is which formula
is the best for which purpose?  For the effect y if exposed to x (eg x is a
treatment), the key CONTRASTing formulas in epidemiology and in EBM ie in
evidence-based medicine for binary  x  are (for multivalued x or for any other
kind of exposure just replace ~x by z ) :

 ARR = P(y|x)  - P(y|~x)  =  Absolute risk reduction aka attributable risk ,
                            "absolute" ie not "relative", often |ARR| too
     = a/(a+b) - c/(c+d)     in a 2x2 contingency table (find 2x2 )
     = [ Pxy - Px*Py ]/[  Px*(1 - Px) ] <= 1  even for tiny Px > 0
     =        cov(x,y)/var(x)  =  slope(of y on x)  =  beta(y:x)   <= 1
     = 0 if x,y are independent (then also P(x|y)=Px, Pxy=Px*Py & P(y|x)=Py)
   !  = 0 to be enforced if Px=1 then P(y|~x)=0/0 & Py=Pxy=Px*Py & P(y|x)=Py
   !  = 0 to be enforced if Py=1 then P(x|~y)=0/0 & Px=Pxy=Px*Py & P(x|y)=Px
    With these enforced values we get a more meaningful ARR (but Py=1 or Px=1
    are too extreme to be of much importance). For DISCOUNTing of the lack
  ! of surprise in y, the measure  P(y|x) - Py <= 1-Py  is better (find SIC )
  since too common y is seldom perceived as much of a risk anyway (find RDS )
 ARR == PNS under exogeneity=no-confounding & monotonicity=no-prevention by
            exposure to the risk factor { Pearl 2000, pp.289,291,300 } ;
       PNS = Probability of Necessity and Sufficiency (in general).

 NNT = 1/|ARR|  = Number needed to treat for 1 more |or 1 less| good effect y
 NNH = 1/|ARR|  = Number needed to harm 1 more |or 1 less|  by side effects z
 NNS = 1/|ARR|  = Number needed to screen to find 1 more |or 1 less| case
 NNE = 1/|ARR|  = Number needed for 1 extra effect { Feinstein 2001, p.172 }
      1/|ARR| is the most realistic measure of health effects in general, as
!!!            it is the least abstract & least exaggerating ie most HONEST,
!!!            and moreover NNT, NNS also measure EFFORT PER EFFECT.
 NNH(z:x)/NNT(y:x)  is also highly informative. It should be >> 1 ie many
                    more have to be x-treated before 1  z-harm will occur,
                        while many more patients have  y-improved already.

 OR  = Odds ratio  =  LR+/LR-  (far below find more on Odds , LR- ).
 LR- = P(~x|y)/P(~x|~y) = LR- = negative LR = ( 1 - sensitivity )/specificity
 LR  = P( x|y)/P( x|~y) = LR+ = likelihood ratio = sensitivity/(1-specificity)
```

```
                             = simple Bayes factor B(x:y)
 RR  = P( y|x)/P( y|~x) = relative risk = risk ratio (unlike ARR, NNT, NNH,
                          RR(:) seems more "impressive" to the innocents) ;
                          RR(:) is a part of important, meaningful formulas :
 PFR = -ARR/P(y|~x) = 1 - RR = Prevented fraction      =  -RRR :
 RRR =  ARR/P(y|~x) = RR - 1 = Relative risk reduction = Excess relative risk
 ARP =  ARR/P(y|x ) = RRR/RR =  1 - 1/RR = Attributable risk percent =
                                         = Attributable risk for exposed =
     = Attributable proportion = etiologic fraction for exposed group = EFE =
     = Attributable fraction in exposure group = AFE   =
     = Excess risk ratio = ERR  { Pearl 2000, p.292 } Since to err is a word,
                           ERR is cannot be found on www :-( My abbreviations
       ARP = EFE = AFE, RDS, PFR and PRP are chosen as findable non-words.

 RDS = ARR/P(~y|~x) =  ARR/[ 1 - P(y|~x) ] = relative difference a la Sheps
     = [ P(y|x) - P(y|~x)]/[ 1 - P(y|~x) ]   for binary  x ; for non-binary :
     =   slope(of y on x) /[ FICTIVE Max. slope of y on x ]   (find as-if )
 RDS = [ P(y|x) - P(y|z) ]/[ 1 - P(y| z) ] = relative difference by Sheps
     = [  successful y if x  minus if z  ] / [ failure rate of y if z ],
    as  P(y|x) <= 1=MAX, the 1 - P(y| z) is the MAXimal thinkable value of
    !   RDS's numerator, ie  1 - P(y| z) is a meaningful normalization. Also
    !!  the IDEA is that failures if z are available to become successes if x
    !!  and that RDS is more honest than RRR, ARP ie AFE if P's are small,
        as they often are, which inflates the latter measures. M.C. Sheps'
        RDS of 1958 { Feinstein 2002 p.174 } can be found for  z == ~x  in :
    - { Patricia Cheng 1997 }  as eq.(16) = RDS,    eq.(30) = 1 - RR = -RRR
      { Novick & Cheng 2004 } too ;
    - { Pearl 2000 } on pp. 284, 292 : PS = RDS,   PN = ERR = 1 - 1/RR = ATE
 but think! :
    P(y|x) - P(y|~x) = ARR is Absolute risk reduction of the effect y , but
    P(x|y)     measures of how much  y implies x  ie y Suffices for y, hence :
!!  P(x|y) - P(x|~y) is a measure of y --> x  or how much is x Necessary for y

!!  P(x|y) - Px  <=  1-Px  DISCOUNTS the lack of SURPRISE in x     (find SIC )
                           since a common  x  is not seen as a real CAUSE ;
      if Pxy=Py then P(x|y) - Px = 1 - Px
      if Pxy=Px then P(x|y) - Px = Pxy*(1/Py - 1) = P(x|y)*(1 - Py)
                                 [ = Px *(1/Py - 1) ]     <=  1 - Px (find SIC )
    where [.] may suggest that Py, Px can be varied, but Pxy <= min(Px, Py).

!!  P(y|x) - Py  <=  1-Py  DISCOUNTS the lack of SURPRISE in y     (find SIC )
                           since a common  y  is not seen as a real RISK ??
      if Pxy=Px then P(y|x) - Py = 1 - Py
      if Pxy=Py then P(y|x) - Py = Pxy*(1/Px - 1) = P(y|x)*(1 - Px)
                                 [ =  Py*(1/Px - 1) ]     <=  1 - Py (find SIC )

 PRP = Pep*(RR-1)/[ 1 + Pep*(RR-1) ]   where  Pep = P(exposed in population)
     = Population attributable risk percent   ( RR is for the studied group)
     = population attributable fraction = etiologic fraction for community

 F(y:x) = ARR/[ P(y|x) + P(y|~x) ] = (RR -1)/(RR +1)  by { Kemeny 1952 }
        = -F(y:~x)   and anaLogically for any mix of events x,y,~y,~x since
                     An event is an event is an event (sorry Gertrude :-)

!! Health effects can be expressed either in negative terms (eg ill, or dead)
   or in positive terms (cured, or alive). Hence we are free to replace any
P(y|.) with P(~y|.) = 1 - P(y|.), in any formula, consistently of course. As
P(.|.)'s are often quite small, 1 - P(.|.) =. 1.  The results will be then
very different, depending on our choice of +terms or -terms. These facts
create ample opportunities for honesty/dishonesty, for leading/misleading.
Clearly, if P(y|~x) < 0.5 then RDS < RRR which only seems more "impressive".
Honestly IF P(y|~x) < 0.5 THEN RDS should be used ELSE RRR should be used.
```

```
        IF P(y|~x) =. 0  THEN RDS =. ARR
```

Of course ARR <= RDS, so ARR can never mathematically exaggerate an effect.

.- +Key construction principles of good association measures :

P1: "Measures of association should have operationally meaningful interpretations that are relevant in the contexts of empirical investigations in which measures are used." { Goodman & Kruskal, 1963, p.311, there also in the footnote } Henceforth I discuss events x, y, but it all holds for their expected values ie averages over variables X, Y ie sets of events too.

P2: OpeRational meaningfulness is greatly enhanced if a measure has its range of values with 3 fixed points of fixed meanings, eg [0..1..oo] or [-1..0..1], where the midpoint means independence, and the endpoints mean extreme dependence (-....+), ideally an implication aka entailment. Yet there are arguments for the range [-Px..0..1-Px] (find SIC Kahre).

P3: Various results from a single measure should be meaningfully COMPARABLE regardless of the total count N of all joint events in a contingency table. This means that a measure should be built from proportions P(:) only, without an uncancelled N. Thus measures based on ChiSquare do not qualify for our purposes. But N must play role in confidence intervals.

P4: To measure association means to measure statistical dependence. I can list 16+1 = 17 equivalent conditions of independence ie equalities lhs = rhs, like eg Pxy = Px*Py, or P(y|x) = P(y|~x), from which 2*17 = 34 measures of dependence can be made by CONTRASTing: lhs - rhs like ARR(:), or lhs/rhs like RR(:) above, both asymmetrical wrt events x, y. Eg the

```
   Pxy/(Px*Py) = P(x|y)/Px = P(y|x)/Py  is symmetrical wrt  x, y, and the
           correlation coefficient is also symmetrical wrt  x, y :
   Sqrt(r2) = Sqrt[ beta(y:x) * beta(x:y) ]
            = Sqrt[ (slope of y on x) * (slope of x on y) ]
                note that -1 <= beta <= 1;  find  r2  below.
```

Measures of confirmation, evidence, indication, influence,.., and of course causation should be DIRECTED ie ORIENTED ie ASYMMETRICAL wrt events x,y. Asymmetry is easily obtained by taking a symmetrical association measure (lhs - rhs) and dividing it by either lhs or rhs, or by 1 - rhs, or by normalization with a function of one variable only, eg:

```
  ARR(y:x) = (Pxy - Px*Py)/(Px*(1-Px)) = cov(x,y)/var(x) = beta(y:x)
           =  P(y|x) - P(y|~x)
```

P5: Measures of CAUSATION tendency should be decomposable into a product of terms such that one term itself measures probabilistic IMPLICATION ie ENTAILMENT, but the equality Measure(y:x) = Measure(~x:~y) is UNDESIRABLE. Alas, the conviction measure conv(y --> x) = conv(~x --> ~y) by Google's CEO { Brin et al 1997 } does not qualify (find UNDESIRABLE). Entailment provides a link with the notions of necessity and sufficiency where (y implies x) == (y is Sufficient for x) == (x is Necessary for y).

P6: Measure(y:x) should yield meaningful values if Pxy = 0; and if Px = 1 :

```
  eg:  RR(y:x) = 0 if Pxy = 0  ie if x,y are disjoint events
  !    RR(y:x) = 1 if Px  = 1  hence  Py - Pxy = 0  AND YET Pxy = Px*Py,
                 1 means independent x,y [ find Pxy/(0/0) as special case ]

   conv(y --> x) = Py*P(~x)/P(y,~x) = [ Py - Px*Py ]/[ Py - Pxy ] in general;
                 =    P(~x)/P(~x|y) = [  1 - Px    ]/[  1 - P(x|y) ]
                 = 0/0 numerically if Px = 1  whence   Py = Pxy   hence:
                 = 1  if Pxy=Px*Py also [ Py - 1*Py]/[ Py - Py ] = 1 algebra
  !!        = 1 - Px  if Pxy=0;  this is not a nice fixed value, but
              1 - Px  is interpretable as "semantic information content" SIC
```

```
                    which makes NO SENSE for Pxy=0 :-( , nevertheless  :-):
      1 - Px < 1 = for x,y independent, so for Pxy=0 is conv < neutral 1 :-)
  !!   Similarly P(~y|~x) = 1-Py makes NO SENSE for Pxy = Px*Py,
             if P(~y|~x) is  ~y Necessary for ~x  ie x Necessary for y

  To avoid overflows due to /0, such extreme/degenerated/special cases of P's
  must be numerically prechecked and detected at run time and handled apart
  according to the meaningful interpretation (or conventions) as just shown.
!! Since any single formula is doomed to measure a mix of at least 2 key prop-
  erties ( dependence and implication mixed due to my INDEPendent-IMPlication
  PARADOX ), it is a good idea to detect & report important extreme/special
  cases which do not always obviously follow from the values returned. Such
! automated reporting adds semantics and avoids misreading/misinterpretation.

P7:  Although it is useful to consider the values returned by measures under
  extreme circumstances like eg Px=1 or Py=1, these will not occur too often,
  and should be prechecked apart anyway. It is more important to choose a
  measure which will return reasonable values for the application at hand.
  We cannot hope that there ever will be a single universally best measure.

  So far for my key construction principles. More analysis follows.

  RR(y:x) is compared with few related measures like eg:
   W(y:x) = weight of evidence by I.J. Good (Turing's statistical assistant);
   F(y:x) = degree of factual support by John Kemeny ( Einstein's assistant);
   C(y:x) = corroboration by Karl Popper (he often called it confirmation, an
            overloaded term,  so Popper corroborates here to be findable);
            it is funny that Sir Popper who stressed refutation has worked
            out measures of confirmation, but not of refutation :-) Why?
conv(y:x) = conviction conv(y --> x)  by Google's CEO Sergey Brin et al.

Such comparisons increase our insights.  How well these formulas measure
causal tendency is also discussed. All this & much more was/is implemented
in my KnowledgeXplorer program  KX  which not only infers & indicates (ie
identifies, diagnoses, predicts, etc) but also extracts knowledge (on both
event- & variable level of interest) from the information carried by data
input in the simple format.  KX has graphical and numerical outputs in
compact, comparative, hence effective forms (eg my squashed Venn diagrams).


.- +MicroTutorial on key elements of probabilistic logic :

  There are 16+1 = 17 equivalent ==    relations for independence = , and
                 17 for -dependence < , and 17 for  +dependence >  of x,y :
   the ? stands for any single symbol < , = , >  consistently applied :

  [Pxy ? Px*Py] == [P(x|y) ? Px] == [ P(y|x) ? Py] ==... [P(~y|~x) ? P(~y|x)]
eg
  [Pxy - Px*Py  ? 0 ] == [ P(y|x) - Py    ? 0 ] == etc ... 17 times
  [Pxy / Px*Py  ? 1 ] == [ P(y|x) / Py    ? 1 ] == etc ... 17 times
eg
  RR(y:x) = [ P(y|x)/P(y|~x) ? 1 ] == [ P(x|y)/P(x|~y) ? 1 ] = RR(x:y)

  Formulas left of an ? are candidate elements for a measure of (x CAUSES y).
  Other elements for (x CAUSES y) must be derived from logic. For 2 binary
  variables there are 16 different logical functions, of which only the
  2 implications and 2 inhibitions are ASYMMETRIC ie DIRECTED ie ORIENTED
  (the remaining 12 functions are either symmetric wrt x,y, or are functions
   of 1 variable only, either x only or y only). Clearly  (x CAUSES y)  must
  to be ASYMETRICAL wrt x,y. But there are more requirements.

  P(~x,~y) = P(~(x or y)) = 1 - (Px + Py - Pxy)   by P(Occam-DeMorgan's law)
```

```
   ~(y,~x) == (y --> x) == (~x --> ~y) == ~(~x,y) == (x or ~y)      in logic.

  Let y  = the observed effect ie evidence;   x = a hypothesised cause of y :
  P(x|y) = Pxy/Py is a NAIVE measure of how much  y suffices to determine x
  P(x|y) = 1 = max iff y --> x  ie  y implies x deterministically ie Pxy = Py
  P(x|y) = Px      iff x,y are independent ie  Pxy = Px*Py.   In extreme case
        of Px = 1 it holds:  if Px=1 then Py = Pxy = Px*Py  AND  P(x|y) = 1
!! ie  y determines x 100%  AND  x,y are independent (seems PARADOXical).

If Px > 0  &  Py > 0  &  Pxy > 0  then
if Px > Py then P(x|y) > P(y|x)   else
if Px = Py then P(x|y) = P(y|x)   else
if Px < Py then P(x|y) < P(y|x)   else Mission Impossibile.

So far on relatively unproblematic sufficiency; now on less clear measures of
necessity:

Pioneers { Buchanan & Duda 1983, p.191 } explained the rule  y --> x  thus :
"... let P(x|y) denote our revised belief in  x  upon learning that y is true.
... In a typical diagnostic situation, we think of x  as a 'cause' and y as an
'effect' and view the computation of P(x|y) as an inference that the cause x
is present upon observation of the effect y."  (find +Folks' for more).
My preferred wording is:  P(x|y) is a NAIVE measure of how much evidence does
 y provide for x  ie how much y implies x as a potential cause of y, hence
!!!  also how likely x CAUSES y.
!!! Note that in P(x|y)  the y = evidence,  x = a hypothesised cause. Ideally
   x CAUSES y if Pxy = Py ie y implies x ie y --> x , eg if P(x|y) = 1.

Causation assumes that without a cause x there will be no effect y , hence
that a cause x is NECESSARY for effect y which then serves as an evidence
for that cause. From the reasonings on the last dozen of lines few candidate
measures (marked by their +pros, -cons, .neutrals ) follow :

1.  P(x|y) = Pxy/Py  is a NAIVE measure of how much is x necessary for y
          -  is not a fun(Px), eg canNOT discount lack of surprise if Px =. 1
          .  is = Px if Pxy = Px*Py ie  x,y independent
          +  is = 0  if Pxy = 0     ie  x,y disjoint
          +  is = 1  if Pxy = Py    ie  P(y,~x) = 0   from Pyx + P(y,~x) = Py
                     ie "without x no y" ie "x necessary for y" (draw a Venn)
                     ie    "if y then x" ie "y sufficient for x"
!!        -  is = 1  see the CounterExample few lines below (also find SIC ).
!!     -  is a single P(.|.) while all single P's were REFUTED as measures of
          confirmation or corroboration { Popper 1972, chap.X/sect.83/footn.3
          p.270, and Appendix IX, 390-2, 397-8 (4.2) etc }

    P(y|x) = Pxy/Px  is analogical (just swap x with y)
          1  iff "y follows from x" is the phrase in { Popper 1972, p.389 }
          +  is used in simple Bayesian chain products for multiple cues.

!!  CounterExample shows that P(.|.) is not good enough measure of causation:
      Let  x = a hypothesised cause, a conjecture
             y = a widely present symptom, eg 10 fingers on each hand.
    Then P(x|y) =. 1  ie Pxy =. Py since almost all with y are ill. Yet it is
    neither wise to assume that  y is sufficient for x,
        nor wise to assume that  x is necessary for y.  (find SIC )

2. An alternative single P-measure of how much is x necessary for y :

  P(~y|~x) = [1 -(Px + Py - Pxy)]/[1 - Px] = 1 - ([Py - Pxy]/[1 - Px])
          +  is a function of Px,  Py,  Pxy , but:
         ?-  is 1-Py if Pxy = Px*Py ie  x,y independent; note that
```

```
              1-Py is a measure of "semantic information content" SIC;
       ? does 1-Py make sense if x,y are independent ? I dont think so.
       ? (similar NO SENSE is conv(y --> x) = 1-Px if Pxy=0  ie disjoint)
       ?.  is  = 0 if   1 = Px + Py - Pxy        ( unlikely to occur ? )
        -  is just a single P which all were REFUTED by Popper
        -  is <> 0 if Pxy = 0  ie  x,y disjoint :-( but it can be forced:
                   if Pxy = 0 then NecessityOf(x for y) = 0 ELSE = P(~y|~x)
        +  is  = 1 if Pxy = Py , as explained next :
        +  is  = 1 if y --> x 100%  ie  if y implies x fully  then :

!  Pxy = Py  AND  P(~y|~x) = 1 = P(x|y) ,     which is consistent with logic:

  ~(y,~x) == (y --> x) == (~x --> ~y) == ~(~x,y) are all equivalent in logic.
```

P(~y|~x) = 1 is the only nicely interpretable fixed point.
P(~y|~x) as a candidate has arisen from my COUNTERFACTUAL reasoning:
the semantical Necessity of x for y follows from IF no x THEN no y, ie
removed or suppressed x suffices for removed or suppressed y, ie
  ~x implies ~y ie ~x --> ~y. The COUNTERFACTUALity in human terms says:
IF x disappears THEN y will disappear too. For more find +Folks' wisdom.

Only after I worked out P(~y|~x) above, I came across { Hempel 1965 } where
at the very end of his very long and very abstract paper I could decode his
eq.(9.11) as P(~y|~x). He derived it as a "systematic power closely related
to the degree of confirmation, or logical probability"{p.282} via his eq(9.6)
which is in fact 1 - P(.) ie SIC. On p.283 the last lines tell us why :
"Range and content of a sentence vary inversely. The more a sentence asserts,
 the smaller the variety of its possible realizations, and conversely."(SIC)
"The theory of Range" is a section in { Popper 1972, sect.72/p.212-213 } where
on p.213 Popper refers the notion of [semantic] Range to { Waismann: Logische
Analyse des Wahrscheinlichkeitsbegriffes, Erkenntnis 1, 1930, p.128f. } .

```
3. -Px  <= [ P(x|y) - Px ] <=  1 - Px           { Kahre 2002, p.118-119 }
   -Px if Pxy = 0              1 - Px if P(x|y) = 1 ie Pxy=Py , find SIC
    Note that (1-Px) - (-Px) = 1 ie the absolute magnitudes of both bounds are
    COMPLEMENTary. This makes sense since a REFUTATION of a conjecture means
    CONFIRMation of its COMPLEMENTary conjecture. Yet users like fixed points.

4. Better measures of sufficiency and necessity are RR(:)'s or LR(:)'s, like
   I.J. Good's
       Qnec = P( e| h)/P( e|~h) = RR( e| h) = Lsuf , see Folk1 ;
       Qsuf = P(~e|~h)/P(~e| h) = RR(~e|~h) = [1 - P(e|~h)]/[1 - P(e|h)]
            = 1/Lnec

   Find +Folks' wisdoms for more. These ratios of ratios have ranges with
   3 semantically fixed values, which enhance opeRational interpretability,
   and are not just single  P's  all REFUTED as measures of confirmation or
   corroboration in { Popper 1972, chap.X/sect.83/footn.3/p.270, and in
   Appendix IX, pp.390-392 etc }.
```

.- end of microtutorial .

Deeper insights into RR, LR, and into confounding are gained by dissecting
 RR(:) and LR(:) thus:

```
 RR(y:x) =   P(y|x)/P(y|~x)  ;    y is effect, x is the hypothesized cause,
                                                eg x is exposure or test result
         =   P(y,x)/P(y,~x)           * ( 1 - Px)/Px
         = [ P(y,x)/(Py  -  P(y,x)) ] * ( 1 - Px)/Px   , find " confound "
         =   P(y,x)*( y implies x )   * SurpriseBy(x)
         =   P(y,x)/P(y,~x)           * SurpriseBy(x)
         =      LikelyThanNot(y:x)    * SurpriseBy(x) ;  note that :
```

```
1/P(y,~x) = 1/(Py - Pyx),  or   1 - P(y,~x) = 1 - (Py - Pyx), are measures of
                                  how likely ( y implies  x) ie IF y THEN x ;
         recall that ~(y,~x) == (y --> x) == (~x --> ~y) == ~(~x,y) ;
         note that              Py - Pxy  = P(~x)  - P(~x,~y)  in general ie
     also for imperfect implication       = (1-Px) - [1-(Px+Py-Pxy)] = Py-Pxy
!!!  but equality of fun(y:x) = fun(~x:~y) is UNDESIRABLE for a measure of
     causal tendency (find UNDESIRABLE below to find out why? ).
     Another fun(y:x) is P(x|y) which also measures (y implies x), however :
      100% implication [ P(x|y) = 1 ] = [ P(~y|~x) = 1 ], while for less than
      100% implication   P(x|y)      <>   P(~y|~x)        in general since :
                         Pxy/Py      <>  [ 1 - (Px + Py - Pxy) ]/[ 1 - Px ]
                         Pxy/Py      <>    1 -     ( Py - Pxy)/[   1 - Px ]
      where by DeMorgan's rule  P(~x,~y) = 1 - (Px + Py - Pxy) = P(~(x or y))

!! (y sufficient for x)  ie    (y implies x)  ie:
   (x necessary  for y)  ie  potentially (x CAUSES y), because removal,
     blocking or reduction of ANY SINGLE necessity (out of several required)
    x necessary for y , will annul or suppress its consequent effect y. Draw
    x enclosing y in a Venn diagram, and see that it is necessary to hit
    x to have any chance of hitting the enclosed y, but not vice versa.
      Hence it is the necessary condition which should be seen as a potential
      cause, removal/suppression of which will remove/suppress the effect y.

1/P(x,~y) = 1/(Px - Pxy),  or   1 - P(x,~y) = 1 - (Px - Pxy), are measures of
                                  how likely ( x implies  y) ie IF x THEN y ;
     recall that ~(x,~y) == (x implies y) == (~y implies ~x) == ~(~y,x) ;

LR = P(x|y)/P(x|~y) = RR(x:y)
   =   P(x,y)/P(x,~y)            * ( 1 - Py)/Py
   = [ P(x,y)/(Px  -  P(x,y)) ] * ( 1 - Py)/Py
   =   P(x,y)*( x implies y )   * SurpriseBy(y)
   =   P(x,y)/P(x,~y)            * SurpriseBy(y)
   =      LikelyThanNot(x:y)     * SurpriseBy(y)

From RR, LR and also from a Venn diagram, it follows that since the joint
P(y,x) = P(x,y), it must be only the unequal marginal probabilities Py, Px,
which decide whether (y implies x) more or less than (x implies y) by the
     rule:
!!!  if Py < Px then RR(y:x) >= RR(x:y) ie LR(x:y),
     if Py > Px then RR(y:x) <= RR(x:y) ie LR(x:y), where
                         the = occurs for x,y independent ie RR(:) = 1,
or if Pxy=0=RR(:), as my program Acaus3 asserts. For more find Py < Px below.

IF LikelyThanNot(y:x) < 1   ie   Pyx < P(y,~x)  ie Less likely than not,
   AND  SurpriseBy(x) is large enough  ie Px is low enough
THEN RR(y:x) > 1 may still result due to low Px.
IF   RR(y:x) > 1  AND  LikelyThanNot(y:x) > 1 ie More likely than not ie
      P(x|y) > 1/2 ie  Pxy > Py/2  ie Pyx > P(y,~x) = Py - Pxy  ie  2Pxy > Py
THEN there is a stronger reason for the conjecture that (y implies x) ie that
     (x causes y), than it is if Pyx < P(y,~x) AND RR(y:x) > 1.
The P(x|y) > 0.5 has been:
- required as "the critical condition for confirming evidence" in { Rescher
  1958, 1970 pp.78-79, and on p.84 swapped to P(y|x) > 0.5 };
- recommended as a potent (not just potential) Necessity N of exposure x for
  case y : N > 0.5 in { Schield 2002 sect.2.3 & Appendix };
- considered in { Hesse 1975, p.81 } but dismissed as a single measure of
  "confirming evidence" because  P(x|y) > 1/2 "may be satisfied even if y has
  decreased the confirmation of x below its initial value in which case y has
  disconfirmed x". Mary Hesse (Oxford) then opted for P(x|y) > Px as the
  condition for "y confirms x" aka Carnap's "positive relevance criterion".
```

```
A PARADOXical behaviour of RR(:), and of other formulas, nearby some extreme
     values is identified :
!!!  huge, even infinite RR(y:x) = oo is possible while y, x are almost
     independent !!!

Let:
 == is equivalence ;   rel is >=< ie  < , = , > , etc
 == is IF [.] THEN [_] and vice versa, ie simultaneously IF [_] THEN [.] .

Keep in mind that there are at least 17 equivalent (in)dependence relations:

    [  P(y,x)  rel Py*Px   ]   , which divided by Px or by Py yields :
 == [  P(y|x)  rel Py      ] == [  P(x|y)  rel Px      ]
 == [  P(y|x)  rel P(y|~x) ] == [  P(x|y)  rel P(x|~y) ]
 == [ P(~y|~x) rel P(~y|x) ] == [ P(~x|~y) rel P(~x|y) ] , etc.

Since both relative risk RR(:) and odds ratio OR(:) are in use, it is good
to remember their relationships:

     OR(y:x) = OR(x:y) = ad/(bc) = Pxy*P(~x,~y)/[ P(x,~y)*P(~x,y) ]

   [ OR(:) rel 1 ] == [ RR(:) rel 1 ] == [ Pxy rel Px*Py ]
hence
  If OR(:) rel 1 (ie if Pxy rel Px*Py) then  OR(:) rel RR(:), and vice versa,
eg
  If OR(:)  <  1 (ie if Pxy  <  Px*Py) then  OR(:)  <  RR(:) ;
  If OR(:)  >  1 (ie if Pxy  >  Px*Py) then  OR(:)  >  RR(:)

which means that  if OR(:) > 1  then  relative risk RR(:) will be smaller
     than odds ratio OR(:).       E.g. let only the OR(:) = 2.5 be known (eg
from a meta-study, so that  a,b,c,d,n  are not known and  RR(:) is not
available). Then we may speculate about the corresponding RR(:) > 1 thus:

OR(:) =        ad/(bc)      =. eg     (25*30) /(10*30)      =  2.5 > 1,
                               or     (25*30) /(30*10)      =  2.5    ;
RR(:) = [a/(a+b)]/[c/(c+d)] =. eg [25/(25+10)]/[30/(30+30)] =  1.4 > 1,
                               or [25/(25+30)]/[10/(10+30)] =  1.8    , etc;
For 1 < RR(:)  <  OR(:)  there is less risk than  OR(:)  suggests. Hence we
convert:
!!!     RR(y:x) = OR(y:x)/[ 1 + (OR(y:x) -1)*P(y|~x)]  where If P(y|~x) may
be a guesstimate.

Keep in mind that swapping rows and/or columns in a 2x2 contingency table
may change OR into 1/OR, but RR(:) will always change, in general.


.- !!! +The simplest thinkable necessary condition for CONFOUNDING :

Lets make search for & research of confounders easier & less expensive.

 RR(y:x) = P(y|x)/P(y|~x) ,  y is the effect, x is the hypothesized cause,
                          eg x is exposure, or treatment, or test result.
Lets consider  c  as a competing (against x) candidate cause of y.  Clearly

 RR(y:c) > RR(y:x)  is a necessary (but generally not sufficient) condition
                    for  c to be, rather than  x , a potential cause of  y.
Less natural is the following condition by Jerome Cornfield et al. of 1959
{ reproduced in the Appendix of Schield 1999 } :

 RR(c:x) > RR(y:x)  is necessary for c, rather than x, to be a cause of y.
```

My decomposition of these RR(:)'s into :

```
 RR(c:x) = [ Pcx/(Pc - Pcx) ] * ( 1 - Px)/Px
 RR(y:x) = [ Pyx/(Py - Pyx) ] * ( 1 - Px)/Px
```

readily suggests that (1 - Px)/Px can be dropped from Cornfield's inequality,
ie

```
ie
   [ Pcx/(Pc - Pcx)    ]  >  [ Pyx/(Py - Pyx)    ]
ie
   [ Pcx*Py  - Pcx*Pyx ]  >  [ Pyx*Pc  - Pyx*Pcx ]  which simplifies to:

!!!  P(x|c) > P(x|y)   my simplest necessary condition for  c  overrulling  x
!!!  P(x|c) - P(x|y)   my simplest necessary absolute boost Ab > 0 needed
!!!  P(x|c) / P(x|y)   my simplest necessary relative boost Rb > 1 needed

   [ P(x|c) = P(c|x)*Px/Pc ] > [ P(y|x)*Px/Py = P(x|y) ]  by Bayes rule ;
!!!          P(c|x)/Pc      >   P(y|x)/Py       my Bayesian boost condition
!!!          P(c|x)         >   P(y|x)*Pc/Py    2nd form of necessary cond.
             P(y|x)         <   P(c|x)*Py/Pc    3rd form of necessary cond.
lead to measures:
!!!   P(c|x)/Pc  -  P(y|x)/Py   = ABb(c:x; y:x)  my absolute Bayesian boost
!!! [ P(c|x)/Pc ]/[ P(y|x)/Py ] = RBb(c:x; y:x)  my relative Bayesian boost
!!! [ P(c|x)/Pc  -  P(y|x)/Py ]/[ P(c|x)/Pc + P(y|x)/Py ]    is my absolute
      Bayesian boost kemenyzed to the range [-1..0..1]
```

If abs.boost < 0 or rel.boost < 1
then confounder c CANNOT replace x as a potential cause of the effect y;
ie abs.boost < 0 or rel.boost < 1 SUFFICE to REFUTE c as a competitor with
  x for a cause of y . This is Popperian refutationalism opeRationalized;
see +Mottos for McGinn on Popper , and find the last Spinoza below.

If abs.boost > 0   or rel.boost > 1
then confounder c   MIGHT replace x as a potential cause of the effect y,
but abs.boost > 0 or rel.boost > 1 are only necessary (but not sufficient)
  conditions for c to replace x as a potential cause of y.

Below find Bailey to read that a "globally" collected P(x|y) is more stable
than P(y|x), which can be estimated from a locally collected P(x|y) thus :

```
     P(y|x) = ( P(x|y)*Py )/[ P(x|y)*Py + P(x|~y)*(1 - Py)     ]  by Bayes,

            =  1/[ 1 + P(x|~y)/P(x|y) *  (1 - Py)/Py  ]
            =  1/[ 1 +   ( 1/LR+ )    * SurpriseBy(y) ]
            =  1/[ 1 +  SurpriseBy(y) / LR+           ]    where
```

 Py has to be the proportion of the effect y in POPULATION. Now it is clear
   how much better it is to use my condition P(x|c) > P(x|y) for confounding.

Combining both necessary conditions for c to overrule x yields :

```
!!   RR(y:x) < mini[ RR(c:x) , RR(y:c) ]  is necessary for c, rather than x,
                                                 to be a potential cause of y;
!!!   P(x|y) < P(x|c)  AND  RR(y:x) < RR(y:c)  is its simpler equivalent.
```

Note that the user does not have to evaluate all (remaining) subconditions
after any single one of them is found to be violated, so that c becomes an
implausible competitor of x for potential causation of y.

My new necessary condition above can also be derived from the fact that in

```
  RR(y:x) < RR(c:x)  ie in  P(y|x)/P(y|~x) < P(c|x)/P(c|~x)
```

conditionings |. are the same on both sides of the < , hence the conditional P(.|.)'s can be turned into joint P(.,.)'s since the conditionings annul.

In the { Encyclopedia of Statistics, Update volume 1 , on Cornfield's lemma, pp.163-4 } J.L. Gastwirth's exact condition for (non)confounding is shown. Let me write it in a clearer notation and then simplify it a bit:

```
 RR(c:x)   =  RR(y:x)  +  (RR(y:x)-1)/[ (RR(y:c)-1)*P(c|~x) ]

 RR(c:x)   >  RR(y:x)  is necessary for c, rather than x, to be a cause of y;
                       is Cornfield's necessary (but insufficient) condition.
```

From Gastwirth's equality follows my more concise sufficient condition for c , rather than x , to cause y :

```
 RR(c:x)-1 >  RR(y:x)-1        + (RR(y:x)-1)/[ (RR(y:c)-1)*P(c|~x) ]
 RR(c:x)-1 > [RR(y:x)-1]  *    (           1/[ (RR(y:c)-1)*P(c|~x) ] )
[RR(c:x)-1]/[ RR(y:x)-1]  >  1 +           1/[ (RR(y:c)-1)*P(c|~x) ]
                      lhs > rhs
lhs - rhs ;    (lhs - rhs)/(lhs + rhs)  has a kemenyzed range [-1..0..1].
```

When the reading gets tough, the tough get reading. This epaper has one thing common with an aircraft carrier: there are multiple cabels to hook on and so to land safely on the deck of Knowledge. There is no safety without some redundancy at critical or remote points.

-.-

+Executive summary :

One good picture or example tells more than 10k words, but 1 formula captures infinitely many examples (remember Pythagoras?). The table, without P(|), CI, and RR(:), is from the handbook on evidence-based medicine aka EBM { Sackett 2000, p.77 }, but it could be economical, investment, or other data as well :

Data:	Cases counted			Information extracted by Jan Hajek :		
Cue	y=bad	~y=good	LR	Probab.	Risk ratio	95% Confidence
xi	n(y,xi)	n(~y,xi)	(xi:y)	P(y|xi)	RR(y:xi)	interval CI(RR)
x1: < 15	474	20	51.9	0.96	5.9	5.34 to 6.52
x2: 15-34	175	79	4.8	0.69	2.5	2.25 to 2.78
x3: 35-64	82	171	1	0.32	1 =independ.	0.91 to 1.10
x4: 65-94	30	168	0.39	0.15	0.5	exercise
x5: > 94	48	1332	0.08	0.03	0.05	exercise

```
Sums:   n(y)=809 + 1770=n(~y)
                2570 = n = sum total
P( y) = n(y)/n = 809/2570 = 0.31 = prevalence = prior ie pre-test probability
P(~y) = 1 - P(y)          = 0.69

Sum_i:[ P(y|xi) ] <> 1
Sum_i:[ P(xi|y) ] = (Sum_i:[ P(xi,y)])/P(y) = P(y)/P(y) = 1.
```

Task: From the left half of the 5x2 contingency table of coincidence counts extract information with opeRationally useful interpretations :

```
 P(y|xi) = predictivity  aka post-test probability of a bad outcome
RR(y:xi) = P(y|xi)/P(y|~xi) = relative risk aka risk ratio of a bad outcome
LR(xi:y) = P(xi|y)/P(xi|~y) = likelihood ratio aka simple Bayes factor.
```

Note that an another way to evaluate the above data would be to contrast a line against another line. That would yield at least 5*(5-1)/2 = 10 pairs of data-lines, each pair forming a 2x2 contingency table for which RR(y:xi), LR(xi:y) and CI(.) would be computed. The number of pairs could be doubled by swapping the 2 lines in each pair with different RR(:), LR(:) and CI(.), because unlike odds ratio OR(:), the RR(:) and LR(:) are not invariant under swaps or transpositions, but they have more of opeRationaly meaningful and useful interpretations which OR(:) does not always have: relative risk.

The cue variable X could be discrete (eg binary ie dichotomous), or it can be a continuous X split into 2 or more levels ie intervals. Here it is a diagnostic test with 5 subintervals < 15,.., > 94, which are relevant for use, but once judiciously chosen they stay fixed, and only the collected classification counts matter. Finer partitioning (= quantization aka discretization) of the continuous cue X into more subintervals xi would decrease the joint counts n(y,xi), n(~y,xi) and thus degrade the robustness of all results.

The solution:

```
P(y|xi)  = P(y,xi)/P(xi) = P(y,xi)/( P(xi,y) + P(xi,~y) )
                         = n(y,xi)/( n(xi,y) + n(xi,~y) )
                   eg:   =     474/(474 +20) = 0.96   or  96%

P(~y|xi) = 1 - P(y|xi)  eg: = 0.04 or 4% is the predictivity of good outcome
         = 20/(474 + 20) ; swapping the columns (or meanings) in the table
   would turn risk ratio RR into my WinRatio WR = RR(~y:x) eg for
   optimistic investors :-)

LR(xi:y) =     P(xi|y)       /   P(xi|~y)
         = [ n(y,xi)/n(y)  ] / [ n(xi,~y)/n(~y) ]
     eg  = [    474/809    ] / [       20/1770  ] = 51.9
```

Not only is LR a bit easier to compute (from the data above) than RR, but in medical applications LR will be more stable than RR. LR's can be collected "globally" (eg on national scale) and via Bayes rule (find here below, or use the nomogram at WWW.CEBM.NET Oxford) applied to the individual cases subject to the local prevalence P(y), or applied to the individual prior probability P(y), to obtain what we really want: the post-test probability P(y|x). Find Bayes and Bailey below. It has been pointed out to me by prof. Brian Haynes (McMaster University, Canada) and by prof. Paul Glasziou (Oxford) that it would be misleading to publish P(y|xi), because a physician must use his or her internal prior P(y) of an individual patient and update it (eg via the nomogram) by LR(x:y) of the external population, to obtain patient's P(y|x). So although LR may carry a more generally useful (because more robust ie stable) partial information, RR carries information more meaningful finally and individually: the risk ratio ie relative risk, and more (read on, pls).

```
RR(y:xi) =   P(y|xi)         /   P(y|~xi)
         = [ P(y,xi)/P(xi) ] / [ P(y,~xi)/( 1 - P(xi) ]
         = [ n(y,xi)/n(xi) ] / [ n(y,~xi)/( n - n(xi) ]
  note that n(y,~xi) = n(y) - n(y,xi);  n(xi) = n(y,xi) + n(~y,xi), hence:
         = [ n(y,xi)/( n(y) - n(y,xi)    )] * [(  n - n(xi)  )/n(xi) ]
         = [       1/((n(y) / n(y,xi))- 1)] * [(  n / n(xi)  ) - 1   ]
   eg:   = [       1/(   809/474      - 1)] * [(2570/(474+20)) - 1   ] = 5.9
   or:
         = [ n(y,xi)/n(xi)    ] * [ (   n - n(xi)  )/(n(y) - n(y,xi)) ]
   eg:   = [     474/(474+20) ] * [ (2570 -(474+20))/( 809 - 474    ) ]
         =       474/ 494       *              2076/335              = 5.9
```

Q: The meaning of P(y|xi) is quite easy to grasp, but what about RR and LR ?
A: Obviously RR(y:xi) is a relative risk as it contrasts the probability of a bad outcome y if xi, against the probability of y if ~xi. That's easy,

operationally meaningful, hence useful. But there are other meanings hidden in RR(:) and similar formulas. In this epaper we shall uncover those hidden meanings or interpretations and properties to obtain fresh insights, eg:

```
     RR(y:xi) = P(y|xi)/P(y|~xi) = P(xi,y)*(y implies xi) * SurpriseBy(xi)
                = P(y,xi)/P(y,~xi) * SurpriseBy(xi)
             = LikelyThanNot(y:xi) * SurpriseBy(xi)

LR = RR(xi:y) = P(xi|y)/P(xi|~y) = P(xi,y)*(xi implies y) * SurpriseBy(y)
                = P(xi,y)/P(xi,~y) * SurpriseBy(y)
             = LikelyThanNot(xi:y) * SurpriseBy(y)
```

One ounce of insight is worth one megaton of hardware. By comparing RR(:) with other formulas we shall see how good it is. Also we shall investigate how well does RR(:) indicate causal tendency, if any. Read on, please.

```
The 95% confidence interval CI of RR(:) completes my info-extraction :
               n( xi) = n(y,xi) + n(~y,xi)  from each row of the data
               n(~xi) = n - n(xi) ;          n(y,~xi) = n(y) - n(y,xi)
SeLn(RR) = standard error of Ln(RR)
         = sqrt[ 1/n(y,xi) - 1/n(xi)    +  1/n(y,~xi) - 1/n(~xi) ]
!! Caution:     if n(y,xi) =.  n(xi)   or    n(y,~xi) =.  n(~xi)
                then SeLn(RR) will be tight even for very small marginal
!! count n(xi) which obviously is UNreliable.

E.g.:  n(x,y) = 3,  n(x) = 3,  n(y) = 54,  n = 75 cases of hart patients :

   RR(y:x) = [n(x,y)/n(x)] / [n(y,~x)/n(~x)] = [3/3]/[(54-3)/(75-3)] = 1.41

  SeLn(RR) = sqrt( 1/3 - 1/3  +  1/(54-3) - 1/(75-3) ) = 0.0756
!!      note that  1/3 - 1/3 = 0 contribution to error from 3/3 = P(y|x)
!!       and even  1/1 - 1/1 = 0  :-((
```

Hence SeLn(RR) has not built-in the wisdom of the old German saying "Einmahl ist keinmahl" ie Once is as-if never. For RR = 1.41 the CI is 1.22 to 1.64 , ie RR will not be outside CI in 95% of trials (to put it simply), hence also RR = 1 (meaning independent x,y ie no relative risk) is not expected in 95% of trials. All seems fine while it is not, since low n(x) is UNRELIABLE.

```
The CI formula for a 95% confidence interval is:
 Ln(CI) spans     (Ln(RR) - 1.96*SeLn(RR)) upto    (Ln(RR) + 1.96*SeLn(RR))
    CI  spans  exp(Ln(RR) - 1.96*SeLn(RR)) upto exp(Ln(RR) + 1.96*SeLn(RR)).

The constant 2.576 is for 99% confidence intervals (are wider),
             1.96  is for 95% confidence intervals (are common),
             1.645 is for 90% confidence intervals (are narrower),
```

which means that eg in 95% of trials with the population counts we shall get a RR(:) value within our CI which is based on much lower sample counts. That's what the books suggest, but as I show above, RR's computed from ratios close to 1 are misleadingly considered as if deserving our confidence :-((

Lets analyze another real example, an ECG test result x at a hart clinic : n(y,x) = 21, n(x) = 74, n(y) = 21, n = 75 patients in total data set, from which my KnowledgeXplorer KX computed and listed (among many other tests and results) :

```
  n(~x) = n    - n(x)   = 75 - 74 = 1  ie all but 1 patient tested had x
n(y,~x) = n(y) - n(y,x) = 21 - 21 = 0  !!
P(y| x) = 21/74 = 0.23
P(x| y) = 21/21 = 1.00
P(y|~x) =  0/1  = 0     !! hence  RR(y:x) = oo ie infinite !!
         also the standard error SeLn(RR) = oo due to 1/n(y,~x) = 1/0 = oo
```

The correlation coefficient between events x,y is :

```
 r = [ P(y,x) - Py*Px ]/[ Py*(1 - Py) * Px*(1 - Px) ] = 0.073   is very low
```

The coefficient of determination (does not exaggerate dependence as r does) :

```
r2 = r*r = beta(y:x)*beta(x:y) = 0.28*0.02 = 0.0056     is even lower
```

This real-world example illustrates that RR(y:x) = oo may obtain for almost independent events x,y . I have not seen a book or a paper telling this !!! PARADOXICAL behavior. We see that 100% implication and near independence are not incompatible. So we can have formulas which have nice opeRational interpretation points with identical meanings [-1..0..1], eg :
degree of factual support F(y:x) by { Kemeny 1952 } which is RR rescaled, and measure of corroboration C(y:x) by { Popper 1972 }. I rewrite both only formally, and find them to have very similar forms. Both yield identical 0.0 in the clean-cut case of 100% independence, but each formula may yield a very different result when less clean-cut ie less extreme situation occurs so they will differ in most common situations. In the last example above we get :

```
F(y:x) = [ P(y|x) - P(y|~x) ] / [ P(y|x) + P(y|~x) ]          {    F-form 1 }
       = [   0.23 -    0    ] / [   0.23 +    0    ] = 1 = 100% implication
       = [    Pxy - Px*Py   ] / [ Pxy + Px*Py - 2*Pxy*Px ]    { my F-form 2 }
       = 0 if  x,y independent
       = 0 if Px = 1  ie unsurprising x , then also  Py = Pxy = Px*Py
                      ie x,y independent AND yet P(x|y) = 1
       = 0 if Py = 1  ie x,y independent AND yet P(y|x) = 1 as Px=Pxy=Px*Py
       = [ P(x|y) - Px ] / [ P(x|y) + Px - 2*Px*P(x|y) ] shows that:
       = 1 iff  P(x|y) = 1  (regardless of Px )
!      = 1 if Px < 1 & Py = Pxy ie (y --> x), also in the extreme case:
           if y = x ie for F(x:x) ie when x implies itself
      = -1 if Pxy = 0 ie x,y disjoint
      = -F(y:~x)
     = [ RR(y:x) - 1 ]/[ RR(y:x) + 1 ]
```

vs the very similarly looking, yet differently behaving :

```
C(y:x) = [   Pxy  - Px*Py   ] / [ Pxy + Px*Py -   Pxy*Px ]    { my C-form 2 }
       = [ P(y|x) - P(y|~x) ] / [ P(y|x) + Py/P(~x) ]
       = [   0.23 -    0    ] / [ 0.23   + 21/1  ] = 0.01 = near independence
       = 0 if  x,y independent
       = 0 if Px = 1  ie unsurprising x , then also  Py = Pxy = Px*Py
                      ie x,y independent AND yet P(x|y) = 1 ,
       = 0 if Py = 1  ie x,y independent AND yet P(y|x) = 1 as Px=Pxy=Px*Py
       = [ P(x|y) - Px ] / [ P(x|y) + Px - Px*P(x|y) ] shows that here:
 <= 1 - Px if P(x|y) = 1  eg C(y:x) = 0 if Px = 1   !!! compare with F(y:x)
 <= 1 - Px if Px < 1 & Py = Pxy ie (y --> x), also in the extreme case:
                 if y = x  ie for C(x:x) ie when x implies itself
    1 - Px  is "semantic information content" of x  ( SIC by Popper's design)
            Note that P(x|y)-Px = 1-Px if P(x|y)=1  ( SIC too w/o /norm   :-)
>= -1 if Pxy = 0 ie x,y disjoint
```

!!! Such formulas, including RR(:) and LR(:) which are just rescaled F(:)'s, may become mixed blessings because they inseparably mix measurements of 2 different properties to which each formula is differently (in)sensitive.

Conclusion: Although a single formula is handy to indicate associations of interest, it cannot be blindly relied upon, especially not when it yields extreme values. Other results from other formulas must be checked alongside. Know thy formulas, and thou shalt suffer no disgrace! This is my paraphrase of the great strategist Sun-Tzu who talked about thy enemies.

-.-

+Mottos :

Great minds discuss ideas, average minds discuss events, small minds discuss people. { Adm. Hyman Rickover, father of the US nuclear navy, whose assistant used to be Charley Martin, author of the best unorthodox cmdr CMFiler which I use to do my epaperwork and for all my file handling, for which I suggested 510 improvements (for MiniTrue only 235 :-) }

Somebody has classified people into three categories:
into the uneducated, who see only disorder;
into the half-educated, who see and follow the rules;
and into the educated, who see and appreciate the exceptions.
The computer clearly belongs to the category of half-educated.
{ Heinz Zemanek, IFIP President 1971-1974 }

Indeed, the thoughtful physician recognizes that each incremental advance in scientific knowledge also unmasks new areas of the unknown that demand resolution. Lewis Thomas has recently written: "The greatest single achievement of science [..] is the discovery that we are profoundly ignorant; we know very little about nature, and we understand even less." ...
We strive for an unscalable summit, destined to be forever obscured in the mists of undiscovered knowledge. ... However our progress is impeded by true ignorance: lack of familiarity with that which is known and lack of comprehension of the need for - and the very nature of - the process of biomedical research. { Thomas H. Weller (1915-????), Nobel prize for medicine 1954, The mountain of the unknown; Hospital Practice, May 1982, pp.33+38+43 }

When you can measure what you are speaking about, and express it in numbers, you know something about it. But when you cannot, your knowledge is of a meagre and unsatisfactory kind.
{ William Thompson aka Lord Kelvin of Larg (1824-1907) }

What is measurable is managable.

In general we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations. ... The proper definition of a concept is not in terms of its properties but in terms of actual operations. ... Meanings are operational.
{ Percy W. Bridgman (1882-1962), Nobelist; father of operationalism 1927 }

Measures of association should have operationally meaningful interpretations that are relevant in the contexts of empirical investigations in which measures are used.
{ Goodman & Kruskal, 1963, p.311, there also in the footnote }

The true logic of this world is the calculus of probabilities.
{ James Clerk Maxwell (1831-1879) }

An event is an event is an event { my paraphrase of Gertrude Stein who thus spoke about a rose. Sorry Gertude, me no Einstein, but Bayes rule and all the independence conditions like P(y|x) = P(y|~x) hold for any mix of x,y,~y,~x }

I don't talk things, sir, said Faber, I talk the meanings of things.
... [Books] have quality. To me it means texture. ... Telling detail.
Fresh detail." { Ray Bradbury, Fahrenheit 451, Part II, pp.75, 83 }

One ounce of insight is worth one megaton of hardware.

Connect, always connect. Compare, always compare. { JH }

God is in the detail. { Mies van der Rohe, architect }

Would that I could discover truth as easily as I uncover falsehood.
{ Cicero }

Detecting error is the primary virtue, not proving truth. ...
There is nothing quite like a brilliant and beautiful theory that has been decisively refuted.
{ Colin McGinn, Looking for a black swan = review of 4 books about/by Karl Popper, in New York Review of Books, November 21, 2002 }

I know you believe you know what I know, but I don't know whether you know what I don't know. { a private thought of every expert (system) }

One man's mechanism is another man's black box { Patrick Suppes, Stanford }

One man's data is another woman's noise;
one man's cause is another woman's effect. { JH }

An invasion of armies can be resisted but not an idea whose time has come.
{ Victor Hugo }

Rerum conoscere causas.

Same cause, same effect { Hempel 1965, p.348 }

Seek simplicity and distrust it. { Alfred North Whitehead (1861-1947), Cambridge, London, Harvard, co-author of Principia Mathematica }

Keep it simple, but not simplistic { Jan Hajek }

Know thy formulas and thou shalt suffer no disgrace { my paraphrase of the greatest strategist ever, Sun-Tzu, 2500 b.PC., : Know thy enemies .. }

Math is hard. Let's go shopping. { Barbie }

-.-

+Combining : priorities, averages, median

Contrasting is done by computing an absolute difference, or a ratio ie a relative difference. An asymmetric denominator provides for the vital ASYMMETRY or ORIENTEDness ie DIRECTEDness. I say: It's the denominator, student!
Combining two semantically different measures (in different units) is generally best done by multiplying them. IF you have to decide between 2 equally available or expensive objects (eg machines) or subjects (eg [wo]men :-) and you have no further information, knowledge, preferences except that the 1st has equally desirable key parameters pA1 and pB1 ,
and the 2nd has equally desirable key parameters pA2 and pB2 ,
where pA's have meaning (eg units of measurement) different from pB's ,
and pA1 and pA2 have the same meaning (eg units) but different values,
and pB1 and pB2 have the same meaning (eg units) but different values,
THEN the golden rule is to buy/choose the object/subject with the larger product pA * pB. E.g.:

```
(S)he1 has IQ1 = 103 and Salary1 = 60k, so that 103*60 = 6180
(S)he2 has IQ2 =  98 and Salary2 = 70k, so that  98*70 = 6860 , then
```

your best pick is (S)he2, ceteris paribus. Feel free to combine IQ with, say the breast/waist ratio, or decide between 2 PC's with different MHz and different GigaBytes for the same price, or freely available from a PC-dump. More math on this (with a threshold value) is in { Grune 1987 }.
Different preferences can be captured by assigning weight wA to paramA and weight wB to paramB, etc. Make sure that weights >= 0, and params >= 1,

because eg (0.4)^2 = 0.16 < 0.4, but we are free to rescale all params so that they all will be >= 1, so there will be no problem. Then the formulas become :

```
 priority1 = (paramA1^wA)*(paramB1^wB)*...etc for more params
 priority2 = (paramA2^wA)*(paramB2^wB)*...etc for more params
 priority3 = (paramA3^wA)*( etc   for 3 objects or subjects to choose from.
```

The maximal priority wins. I have extensively used heuristic priorities computed & pushed in priority queues during my pioneering R&D on automated verification of communication / networking protocols back in 1977 done via the thinktank RAND Corp. for DARPA (find both on WWW), when TCP was still fresh & buggy. See my epaper on APPROVER on WWW.MATHEORY.INFO or .COM

The above task is related to the averages based on multiplication (rather than on addition) :
harmonic average ha = 2*A*B/(A+B) <= sqrt(A*B) = ga = geometric average

which both (unlike the arithmetic average) yield zero when either A or B is zero. For our task above the geometric mean would be fine, while harmonic average would be harmful as it contains (A+B) which makes no sense
! if A and B have different meanings (eg different units). Multiplication makes sense in general, since you dont want someone with IQ = 0 or with the breast/waist ratio = 0, do you ? See www for "weighed averages".

Arithmetic mean minimizes the variance ie the sum of squared deviations from the mean.
The median minimizes the sum of ABSolute deviations from the median.
I cannot go here into further criteria for when to use which kind of average and median, but a good book on statistical literacy should go.

-.-

+Notation, tutorial, basic insights, PARADOXical "independent implication" :

```
 ?(y:x) denotes a measure of how much the evidence y CONFIRMS x as a cause,
       conjecture or hypothesis x. (y:x) means that y implies x  ie y --> x,
       within such a measure. Note that  P(x|y), or P(x|y) - Px are y --> x
       where -Px  discounts the lack of surprise if Px is high (find SIC ),
       while in  RR(y:x) = P(y|x)/P(y|~x) it is the 1/(Py - Pxy) == y --> x
       [ in /P(y|~x) ], which due to its range overrules  P(y|x) == x --> y ,
       while (1-Px)/Px discounts the lack of surprise if Px is high (find SIC)
       but don't get misled by the form, since most measures can be rewritten
       as :
         [ Pxy - Px*Py ]/[ denominator1 ] = cov(x,y)/denominator1
       = [ P(x|y) - Px ]/[ denominator2 ]  if denominator2 = 1 then y --> x
       = [ P(y|x) - Py ]/[ denominator3 ]  if denominator3 = 1 then x --> y

       where the numerator   captures dependence (is 0 if x,y independent),
       while the denominator decides implication y --> x, or x --> y.
       It's the denominator, students! :-)  For example :

       cov(x,y)/Py = P(x|y) - Px
       cov(x,y)/Px = P(y|x) - Py
       cov(x,y)/var(x) = cov(x,y)/[Px*(1-Px)] = P(y|x) - P(y|~x) = ARR
       cov(x,y)/[ Pxy + Px*Py -   Pxy*Px ]    = C(y:x)  { my C-form 2 }
       cov(x,y)/[ Pxy + Px*Py - 2*Pxy*Px ]    = F(y:x)  { my F-form 2 }

       Popper, Kemeny and I.J. Good used ?(x:y), until in 1992 I.J. finally
         switched to my less error prone ?(y:x) which is more mnemonical, as
       it matches the 1st term which is  P(y|x) in their formulas.
```

```
[0..1..oo) is a half open interval including 0 but excluding infinity oo,
           where the central point 1 means stochastic independence. Since I
       assume marginals Px > 0 and Py > 0, many intervals will be half open
       ..oo) and not ..oo] ie not closed.

as-if  useful fiction a la { Vaihinger 1923 }.  E.g. the product of marginal
       probabilities  Px*Py  provides a fictional point of reference for
       dependence of events x and y. Fictional, because  Pxy = Px*Py  occurs
       rarely, but we often contrast  Pxy vs Px*Py   in  Pxy - Px*Py  or in
       log(Pxy/(Px*Py)) in Shannon's mutual information formula. I realized
       that the Archimedean point of reference { Arendt 1959 } may be a
       special case of the useful as-if fictionalism (find Px*Py here & now).
       Also the MAXimal possible values are as-if values eg for normalization.
  SIC  stands for "semantic information content" ( sic! :-) ; find Popper
  <>   unequal ie either smaller < , or greater > then something
   =   equal
   =.  nearly or approximately equal, close to
  :=   assignment statement in Pascal (in C it is the ambiguous sign = )
  ==   equivalence of two terms, or synonymity of notations or terms
       equivalence is a logical relationship R(a,b) such that it is:
          reflexive & symmetric & transitive;
          reflexive(a)    := R(a,a)           for all a
          symmetric(a,b)  := R(a,b) & R(b,a)  for all pairs of a,b
          transitive(a,c) := R(a,b) & R(b,c)
  <=   less or equal
  >=   greater or equal
  >=<  any one of the relations > = <  >=  <=  <>  consistently used
   *   multiplication
  oo   infinity (eg 1/0 = oo , but 0/0 = undefined in general, but in
         expected values we take 0*0/0 = 0 , eg in entropic means; in
         RR(y:x) = conv(y --> x) = 0/0 = 1 if Px = 1 ie Py = Pxy = Px*Py ! )
   ^   power operator, eg  3^4 = 3^2 * 3^2 = 9*9 = 81
  sqr(.) == square(.) == (.)^2 = (.)*(.)
  sqrt(.)      is a square root of (.)
  Sum_i:[ . ]  is a sum over the items indexed by  i  within [.]
  lhs, rhs   abreviate left hand side, right hand side respectively
  exp(a) = e^a  where e = 2.718281828  is Euler's number
      ln(.)         is logarithmus naturalis based on Euler's constant e
  exp(ln(.)) = (.)  is antilogarithm aka antilog
  log2(a)  =  ln(a)/ln(2) = ln(a)/0.69314718 = 1.442695*ln(a) , now base = 2
  log(a*b) = log(a) + log(b)    where log(.) is of any base, eg ln(.)
  log(a/b) = log(a) - log(b) = -log(b/a);    log(1/a) = -log(a)
     (a/b - b/a) = -(b/a - a/b) is a logless reciprocity function of a, b
     (a   - 1/a) = -(1/a - a  ) is a logless reciprocity function of a.
      Reciprocity is desirable when creating new entropy functions.
      A logless additivity can be achieved by relativistic regraduation.

  x, y, e, h   symbolize events    viewed as-if random events    r.e.s
  X, Y         symbolize variables viewed as-if random variables r.v.s,
                                        here an r.v. is a set of r.e.s
 ~x    negation (ie complement) of an event x , so that P(~.) + P(.) = 1
  P(.) is a probability, a proportion or a percentage/100. Empirical P's in
       general and observational P's in particular should be smoothed from
       the range [0..1] to (0..1) ie to 0 < P < 1.
       There are several definitions of probability, the main distinction is
       frequentist (based on repetition and exchangebility) vs subjectivist
       (allowing plausibility or belief). I am an unproblematic guy because
       antidogmatic Bayesian frequentist or a data-driven empirical Bayesian.
       Here I see each proportion as an approximation of a probability.
       In fact a proportion is a maximum likelihood (ML) estimate of a
       probability, which is ok if in  c/n  the count c > 5 and n = large.
       I designed robust formulas for estimates when  c = 0, 1, 2, 3, etc,
       and data-tested their great powers in my KnowledgeXplorer aka KX.
  Px == P(x)  is a parentheses-less notation for P(x_i) ie P(x[i]) ie P(xi).
  1-Px has range [0..1]; it linearly decreases with Px, and it measures:
        + improbability of an event x ;  x may be a success or a failure;
        + surprise value of  x ;  the less probable, the more surprising is
              x  when it happens. What is too common, cannot be surprising.
       What is not surprising is not interesting, carries no new meaning.
       More surprising x means more "semantic information CONTENT in x" SIC ,
       since the lower the Px the more possibilities it FORBIDS, EXCLUDES,
           REFUTES or ELIMINATES when x occurs (find SIC , Spinoza below).
  1/Px has the range [1..oo) and it hyperbolically decreases with Px.
  log(1/Px) = -log(Px)  ranges [0..oo); log bends the steep 1/Px down, and
               measures surprise in Shannon's classical information theory.
     In 1878 Charles Sanders Peirce (1839-1914) has linked log(Px) to Weber-
             Fechner's psychophysical law, see { Norwich 1993 }.
     In 1930ies Harold Jeffreys wrote about log[ LR(x)/LR(y) ], Abraham Wald
     in 1943, and Turing & Good used this "weight of evidence" during WWII.
 (1-Px)/Px  ranges [0..oo);  is my steep measure of surprise in an event x .

  E[f(x)]   = Sum[Px*f(x)] = expected value of f(x) ie an arithmetic average
                             ie an arithmetic mean of f(x). Let f(x) = P(x) :
  E[ Px ]   = Sum[Px*Px] = Sum[Px^2] = expected probability of the variable X
  1 - E[Px] = Sum[Px*(1 - Px)] = Sum[Px - Px*Px] = 1 - Sum[(Px)^2]
            = expected probability of error or failure for r.v. X
            = expected surprise = expected semantic information content SIC
            = quadratic entropy, which is not only simpler and faster than
  Shannon's, but also provably better for classification, identification,
  recognition and diagnostic tasks. Shannon's entropies are better only for
  coding. Don't tell this secret to any classical information theorist :-)
  Variance of an indicator event x (ie binary or Bernoulli event) is:

  Var(x) = Cov(x,x) = P(x,x) - Px*Px = Px - (Px)^2 = Px*(1 -Px),  since
           Cov(x,y) = P(x,y) - Px*Py = covariance of events x,y in general

  Px*Py  is a fictitious joint probability of as-if independent events x, y;
         it serves as an Archimedean point of reference (a la Arendt )
         to measure dependence of x,y  either by  Cov(x,y) = Pxy - Px*Py or
!        by Pxy/(Px*Py), (find as-if ). If Px=1 or Py=1 then Pxy = Px*Py !
  P(x,y) == Pxy == P(x&y) is the joint probability of x&y . Pxy measures
            co-occurrence ie compatibility of x and y.  Until early 1960ies
  P(x,y) had used to denote P(x|y)  in the writings of Hempel, Kemeny, Popper
         and Rescher, while they used P(xy) for the modern P(x,y) ie my Pxy.
         Empirical and observational proportions should be smoothed to :
0 < Pxy < minimum[ Px, Py ] ie an empirical  P(x,y)  should be less than
         its smallest marginal P. Low counts n(x,y) >= 1 are much improved by

  P(x,y) =. [n(x,y) - 0.5]/N ,   and   P(y|x) =. [n(x,y) - 0.5]/n(x)

         which I may show derived exactly (ie = , not just =. ) elsewhere.

  P(x|y)    = Pxy/Py  defines conditional probability, and Bayes rule follows:
  P(x|y)*Py = Pxy = Pyx = Px*P(y|x)     shows invertibility of conditioning
  P(x|y)/Px = P(y|x)/Py = Pxy/(Px*Py)   is my favorite form of basic Bayes as:
  P(x|y) ? Px       ==  P(y|x) ? Py ,     where the ? is < , = , > ; and also
  P(x|y) ? P(x|~y)  ==  P(y|x) ? P(y|~x)  where the ? is applied consistently.
  P(x|y)/P(y|x) = Px/Py                 is Milo Schield's form of basic Bayes
  P(x|y) = Px*P(y|x)/Py                 is the basic Bayes rule of inversion,
     where Px = "base rate"; IGNORING Px is people's "base rate fallacy".
  Odds form of Bayes rule :
  Odds(y|x)        =    Odds(y)  * LR(x:y)  { Odds local or individual,
                                                 LR "global" eg national }
   = P(y|x)/P(~y|x) = (Py/(1-Py)) *  P(x|y)/P(x|~y) = P(y|x)/(1 -P(y|x))
   = P(y,x)/P(~y,x)
   = n(y,x)/n(~y,x) = n(y,x)/[ n(x) - n(y,x) ]  would be the straight, but

     misleading estimate in medicine (find Bailey / Glasziou / Haynes here).

     P(y|x) = Odds(y|x)/(1 + Odds(y|x)) = 1/(1/Odds(y|x) + 1)
            = 1/( 1  + n(x,~y) / n(x,y) )
            = n(x,y)/( n(x,~y) + n(x,y) ) = n(x,y)/n(x) = P(y|x)       q.e.d.

 -log(Bayes rule) :
 -log(   P(x|y)   ) = -log(Pxy/Py) =
 -log(Px*P(y|x)/Py) = -log(Px) - log(P(y|x)) + log(Py)  is the -log(Bayes)

         Note that for only comparative purposes between several hypotheses
    x_j  we may ignore  Py  (but NEVER IGNORE the base rate Px !) since Py is
         a (quasi)constant for all  x_j's compared: the shortest code for max
  P(x_j, y) wins. This holds for logless Bayesian decision-making too: the
            maximal Pxy is the winner. This is Occam's razor opeRationalized,
            as it has the minimal coding interpretation as follows :
    x = unobserved/able  input of a communication channel, or
        unobservable hypothesis/conjecture/cause/MODEL to be inferred/induced;
    y =   observed/able output of a communication channel, or
          available test result/evidence/outcome/DATA.
   According to { Shannon, 1949, Part 9, p.60 } and provable by Kraft's
   inequality, the average length of an efficient ie shortest and still
   uniquely decodable code for a symbol or message  z  will be  -log(P(z)),
   in bits if the base of log(.) is 2.  Hence the interpretations of our
  -logarithmicized Bayes rule { Computer Journal, 1999, no.4 = special
   issue on MML, MDL } are opeRationalized Occam's razors :
    - MML = minimum message length     (by Chris Wallace & Boulton, 1968)
    - MDL = minimum description length (by Jorma Rissanen, 1977)
    - MLE = minimum length encoding    (by Pendault, 1988)
  These themes are very very close to Kolmogorov complexity, originated in the
  US by Ray Solomonoff in 1960, and by Greg Chaitin in 1968, and were designed
  already into Morse code, and by Zipf's law evolved in plain language, eg:
  4-letter words are so short because they are used so often. In Dutch we use
  3-letter words because we either use them more frequently, and/or we are
  more efficient than the Anglos :-))
  Hence the total cost ie length of encoding is the sum of the cost of coding
  the model  x_j , plus the cost ie code size of coding the data  y  given
  that particular model  x_j.  Stated more concisely :

  cost or complexity = log(likelihood) + log(penalty for model's complexity)

       and you know that I dont mean any models on a catwalk :-) The pop
  version of Occam's "Nunquam ponenda est pluralitas sine necesitate" is the
  famous KISS-rule: Keep it simple, student ! :-)  Simplicity should be
  preferred over complexity, subject to "ceteris paribus". Einstein used to
  say: "Everything should be made as simple as possible, but not simpler".

  The MOST SIMPLISTIC, NAIVE measures of causal tendency :

  P(y|x) = Pxy/Px = Sufficiency of x for y  { Schield 2002, Appendix }
                  =   Necessity of y for x  { follows from the next line: }
  P(x|y) = Pxy/Py =   Necessity of x for y  { Schield 2002, Appendix }
                  = Sufficiency of y for x  { follows from above }
    but  WATCH OUT , CAUTION :
!!! let x = a disease,  y = 10 fingers : P(y|x) = 1 in a large subpopulation
    but it would be a semantic NONSENSE to say that  x suffices for y , or
    that y is necessary for x  { courtessy Jan Kahre, private comm. }.
  My analysis: P(y|x) = Pxy/Px is not a DECreasing function of Py, hence any
        y with P(y) =. 1 ie too COMMON y will REFUTE P(y|x)  as a measure.
```

```
!!! Much more complicated REFUTATIONS of all single  P(.|.)'s or P(.)'s as
    measures of confirmation or corroboration are in { Popper 1972, Appendix
    IX, pp.390-2, 397-8 (4.2) etc, and p.270 }. P(.|.)'s should be viewed as
    NAIVE, CRUDE, MOST SIMPLISTIC measures :         rel. = relatively
  P(x|y) = Pxy/Py = a measure of (y implies  x)  ie  rel. how many y are x
                  = a measure of (x includes y)  ie  rel. how many y in  x ;
  P(y|x) = Pxy/Px = a measure of (x implies  y)  ie  rel. how many x are y
                  = a measure of (y includes x)  ie  rel. how many x in  y ;
                  draw a Venn diagram of targets being hit by arrows.
  P(y|x)*P(x|y) =  a measure of (x Sufficient for y) & (x  Necessary for y)
                =  a measure of (y  Necessary for x) & (y Sufficient for x)
                =  ((Pxy)^2)/(Px*Py) ;  its symmetry makes it worthless as
                                     a measure of causal tendency.
  Pxy/(Px*Py) has range [0..1..oo) and measures stochastic dependence;
    oo  unbounded POSitive dependence of x, y
     1  iff independent x, y
     0    bounds  NEGative dependence of x, y
     0  iff disjoint x, y ; do not confuse disjoint with independent !
  A fresh alternative look at old stuff ( Px*Py is as-if independence ) :

  Pxy/(Px*Py) = (Pxy/Px)*(1/Py) =
 =                    (x implies y)*( steepSurprise by y )
 =         (Sufficiency of x for y)*( steepSurprise by y )
 =         (  Necessity of y for x)*( steepSurprise by y )
 = (Pxy/Py)*(1/Px)  = (y implies x)*( steepSurprise by x )
 =         (Sufficiency of y for x)*( steepSurprise by x )
 =         (  Necessity of x for y)*( steepSurprise by x )
 = [0..1]*[1..oo) = [0..1..oo) is the range; 1 iff independent
 = symmetrical wrt x,y  which may be good for coding but poor for a directed
   ie oriented eg causal inferencing, hence I created :
!!
 (Pxy/Px)*(1-Py) = P(y|x)*(1-Py) = (x implies y)*(linearSurprise by y )
 (Pxy/Py)*(1-Px) = P(x|y)*(1-Px) = (y implies x)*(linearSurprise by x )
                 = [0..1]*[0..1] = [0..1] is very reasonable
!         is asymmetrical wrt x, y hence is capturing causal tendency better.
 I created these new measures because trivial ie unsurprising implications
 are of little interest for data miners, doctors, engineers, investors,
 researchers, scientists. The next formulas would overemphasize importance of
 surprise, because Pxy/Px has range [0..1], while (1-Py)/Py has [0..oo) :
!
 (Pxy/Px)*(1-Py)/Py = P(y|x)*(1-Py)/Py = (x implies y)*(bigSurprise by y )
                    = [0..1]*[0..oo)   = [0..oo)      { big range        }
 (Pxy/Py)*(1-Px)/Px = P(x|y)*(1-Px)/Px = (y implies x)*(bigSurprise by x )
                    = [0..1]*[0..oo)   = [0..oo)

 Only after this synthesis we may not be surprised that the last lines are
 a substantial part of a risk ratio aka relative risk :

   RR(y:x) = P(y|x)/P(y|~x)  is 0 for disjoint x,y ; is 1 for independent ;
           = (Pxy/(Py - Pxy))*(1-Px)/Px = (y implies x)*(bigSurpriseBy x)
           =          [0..oo)*[0..oo)   = [0..oo)
 note that :
 +    both factors have the same range [0..oo) hence none of them dominates
           structurally ie in general;
 + in both factors both numerator and denominator are working in the same
           direction for increasing the product of implies * surprise;
 + there is no counter-working within each and among factors.

! P(y|x) > P(y|~x)  ==  P(x|y) > P(x|~y)  ==  Pxy > Px*Py  (derive it) which
                    is symmetrical ie directionless ie not oriented;
  the equivalence holds for the <  <>  =  >=  <= as well, the = is in all
  17 conditions of independence.  On human psychological difficulties in
  dealing with such causal/diagnostic tasks see { Tversky & Kahneman:
  Causal schemas in judgments under uncertainy } in { Kahneman 1982,pp.122-3}

  cov(x,y)  = Pxy - Px*Py = covariance of events x, y (binary aka indicator)
  var(x)    = Pxx - Px*Px = Px*(1 - Px) = variance of an event x  (autocov )
  corr(x,y) = cov(x,y)/sqrt(var(x)*var(y)) = correlation of binary events x,y

  >=    greater or equal.
  => is meaningless in this epaper, although some use it for an implication,
        which is misleading because :
  (y --> x) == (y <== x) == (y subset of x) == (y implies x);    note that
              the <= works on Booleans represented as 0, 1 for False, True
              respectively and evaluated numerically.  E.g. in Pascal
   (y <= x)  on Boolean variables means that  (y implies x).
   In our probabilistic logic ( P(x|y)=1 ) == (y implies x) fully,
      ie (y is Sufficient for x), ie to hit  y will hit  x ,
  !!  ie (x is Necessary  for y), ie to miss x will miss y (just
      draw a Venn diagram with a smaller circle  y  within a larger
      circle  x , ie with full overlap, and view these circles as
      targets to be hit or missed by you, the virtual archer.

  My B(y:x), W(y:x), F(y:x) and C(y:x) have been written as ?(x:y) by ancient
  authors like I.J. Good, John Kemeny and Sir Karl Popper, who were inspired
  by the Odds-forms, which swap x, y via Bayes rule of inversion.  However my
  notation (I.J. Good used it only in his latest papers since 1992) is much
  less error prone as it naturally & mnemonically abbreviates the simplest
  straight forms like eg:

 RR(y:x) = risk ratio = relative risk = B(y:x) = simple Bayes factor
        =  P(y|x) / P(y|~x)

ARR(y:x) =  P(y|x) - P(y|~x)    = absolute risk reduction  = risk difference
                                = attributable risk
        = a/(a+b) - c/(c+d)    = (ad - bc)/[ (a+b)*(c+d) ]
        = (Pxy -Px*Py)/(Px*(1-Px)) = risk increase (or risk reduction )
        =     cov(x,y)/var(x)  =  covariance(x,y)/variance(x)
  !!    =           beta(y:x)  =  the slope of the probabilistic regression
          line Py = beta(y:x)*Px + alpha(y:x) for indication events x, y
          ie for binary events aka Bernoulli events;   -1 <= beta(:) <= 1
   !  0.903 - 0.902 = 0.001 is relatively small, but the same difference:
      0.003 - 0.002 = 0.001 is relatively large; absolute differences may be
      misleading for some purposes, but for practical treatment effects the
 RR(y:x) exaggerates risk more, and more often than ARR(:) and 1/|ARR|'s
        like NNT, NNH do.

RRR(y:x) = RR(y:x) - 1  = ARR(y:x)/P(y|~x)  = [ P(y|x) - P(y|~x) ] / P(y|~x)
        = relative risk reduction
        = excess relative risk = relative effect

  F(y:x) = (P(y|x) - P(y|~x)) / (P(y|x) + P(y|~x)) = factual support
        =      difference    / ( 2*sum/2 )             my 1st interpretation
  !!    =     (difference/2) /  arithmetic average of both P(.|.)'s
        =          deviation /  arithmetic average
  !!    = (slope of y on x ) / (P(y|x) + P(y|~x))      my 2nd interpretation
        =        beta( y:x ) / (P(y|x) + P(y|~x))  ,   -1 <= beta(:) <= 1 ,
        = [ cov(x,y)/var(x)] / (P(y|x) + P(y|~x))
   = (Pxy -Px*Py)/(Px*(1-Px)) / (P(y|x) + P(y|~x))
   = rescaled B(y:x) from   [0..1..oo) to [-1..0..1]      3rd interpretation
   = rescaled W(y:x) from (-oo..0..oo) to [-1..0..1]      4th interpretation
   = is a combined (mixed) measure scaled [-1..0..1] of :
       - how much (y implies x) ,  yielding  +1  iff 100% implication
       - how much  y and x  are independent,  0  iff 100% independence
   = [ ad - bc ]/[ ad + bc + 2ac ]

   = CF2(y:x) = ( P(x|y) - P(x) )/( Px*(1 - P(x|y) + P(x|y)*(1 - Px) )  is

     a certainty factor in MYCIN at Stanford rescaled by D. Heckerman, 1986,
                                 which I recognized to be F(y:x) via my:
   = (Pxy - Px*Py)/(Pxy + Px*Py - 2*Px*Pxy)             my 5th interpretation

   = [ RR(y:x) -1]/[ RR(y:x) +1 ]                       my 6th interpretation

  F0(:)   = [ F(:) + 1 ]/2         is F(:) linearly rescaled to [0..1/2..1] :
  F0(y:x) = P(y|x)/[ P(y|x) + P(y|~x) ]

  all these measures are changing co-monotonically, and they all measure
  - how much the event y implies the x event. This is the directed ie
             oriented ie asymmetric component of these measures;
  - how much x, y are stochastically dependent ie covariate ie associate.
             This is the symmetrical aspect or an association.
  No contortion is needed to have events x, y which are almost independent,
  if we measure independence by Pxy/(Px*Py)  or by (Pxy -Px*Py)/(Pxy +Px*Py)
  or by  (Pxy - Px*Py)/min(Pxy, Px*Py), and at the same time one event will
  strongly imply the other event. But this PARADOX depends on the sensitivity
  (wrt the deviations from exact independence) of the measure. Hence our
  choice of a single measure should depend on our preference for what the
  measure should stress: an implication, or (a deviation from) independence.
  E.g. K. Popper's corroboration C(:) stresses dependence over implication,
  while Kemeny's factual support F(:) stresses implication over dependence,
  but neither of those authors say so, nor anybody has noticed that so far.
  Of course, we could always use two measures, one for an implication, and
  the other for a deviation from independence, but the Holy Grail is a
  single formula, which will inevitably combine ie mix these two aspects,
  because they are almost arbitrarily (but not 100%) mixable.

A disclaimer:  however impossible it may be to find the Excalibur formula
  for causality, I believe it to be possible to identify formulas which come
  closer to the Holy Grail than other formulas. I consider the notions of
  stochastic DEPENDENCE together with probabilistic IMPLICATION (or my
  INHIBITION) and SURPRISE as the key building blocks because they are well
  defined (though not understood enough by too many :-(

A claimer: My goal here is to generate knowledge & understanding of the
  best & the brightest inferrencing formulas for what i call an INDICATION.
  The formulas must provide clear opeRational interpretations, ie they must
  make sense out of the data from which they were computed. There is no lack
  of formulas which somehow capture an association between events. In fact
  there are too many of them, with too many pros & cons.

-.-

+Interpreting a 2x2 contingency table wrt RR(:) = relative risk = risk ratio:

           a , b         the counts  a, d  are hits, and  b, c  are misses
           c , d                 ie  a, d  concord,  and  b, c  discord
                              and  a+b+c+d = n = the total count of events.
!! It is useful to view such a table as a Venn diagram formed by two
   rectangles, one horizontal and one vertical, with partial overlap
   measured by n(x,y) = the joint count ie co-occurrence of x and y :

   ______________________________
   |              |              |
   | a = n( x,y)  | b = n( x,~y) |   n( x) = a+b
   |              |              |
   |-------------|--------------.
   |             |              .
   | c = n(~x,y) | d = n(~x,~y) .   n(~x) = c+d
   |_____________|...............

     a+c =  n(y)   b+d =  n(~y)     N = a+b+c+d
```

but nothing prevents you from viewing the overlap in any of the 4 corners.
Feel free to rotate or to transpose this standard table at your own peril.
Typical semantics (one quadruple per line) may be, eg:

x	~x	y	~y
test+ says K.O.	test- says ok	disorder	not this disorder
exposed	unexposed	illness	not this�illness
risk factor present	risk fact.absent	outcome present	outcome absent
! treatment	control	non-case	case
alleged cause	cause absent	effect	not this effect
symptom present	symptom absent	possible cause	not this cause
conjecture,hypothesis		evidence observed	

so be careful with assigning your own semantics ! We can avoid mistakes
if we stick here to the first four interpretations just listed.
The 2x2 probabilistic contingency table summarizes the dichotomies :

```
      |             y                 ~y  |  marginal sums
 -----|-----------------------------------|-------------------------
    x |  a/n = P( x,y)  ,  b/n = P( x,~y) |  P( x) = (a + b)/n
   ~x |  c/n = P(~x,y)  ,  d/n = P(~x,~y) |  P(~x) = (c + d)/n = i/n
 -----|-----------------------------------|-------------------------
 Sums |           P(y)             P(~y)  |  1 =  P( x,y) +P( x,~y)
      |      = (a+c)/n     = (b+d)/n =f/n |      +P(~x,y) +P(~x,~y)
```

In my squashed Venn diagram in 1D-land, the joint occurrences of (x&y)
ie (x,y) ie "a" are marked by ||| = a/N = Pxy :

```
nnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnnn
ffffffffffffffxxxxxxxxxxxxxxxxxxxxxxxxxxxxxffffffffffffffffffffffffffff
-------------------------aaaaaaaaaaaaaaaaaa----------------------------
iiiiiiiiiiiiiiiiiiiiiiiiiyyyyyyyyyyyyyyyyyyyyyyyyyyyyyyiiiiiiiiiiiiiiii


11111111111111111111 A limited 1-verse of discourse 1111111111111111111
---- 1-Px ----xxxxxxxxxxxxxxxxx Px xxxxxxxx---------------- 1-Px ------
---- 1-Pxy --------------|||||| Pxy |||||||---------------- 1-Pxy -----
---- 1-Py ---------------yyyyyy Py  yyyyyyyyyyyyyyyyy------ 1-Py ------
```

From the 4 counts (a+b+c+d = n) we easily obtain all P(.)'s.
From the 3 proportions or probabilities Px, Py and Pxy we can obtain
any other P(.,.) and P(.|.) containing any mix of (non)negations,
but without raw counts we cannot compute eg confidence interval CI.

The legality of P's (given or generated) can be checked by the following
Bonferroni / Frechet inequalities:

```
Max[     Px , Py         ] <=  P(x or y)  <= min[ 1, Px + Py ]
Max[ 0,  Px + Py - 1     ] <=  Pxy        <= min[    Px , Py ]
```

the lhs of which is the Bonferroni inequality, which becomes nontrivial
only if Px + Py > 1, in which case there will be Pxy > 0.

```
Pxy <= min[ P(x|y) , P(y|x) ]   is my own simple inequality, also useful
                                for checking, and if violated then for
```

```
                                  trimming of eg smoothed estimates.
  The inequality for Pxy divided by Py, or by Px, yields my favorites :

         Max[ 0, (Px + Py - 1)/Py ] <=  P(x|y)     <= min[ Px/Py ,  1 ]
         Max[ 0, (Px + Py - 1)/Px ] <=  P(y|x)     <= min[ Py/Px ,  1 ]

  For the union U of m events  x_i  with probabilities Pi we get :

  the simple          Max_i:[Pi] <= P( U_i:[x_i] ) <= min( 1, Sum_i:[Pi] )

  and if we know P(j,k) ie Pjk  of all pairs of joint events then:

 Sum_i:[Pi] - SumSum_j<k:[ Pjk ] <= P( U_i:[x_i] ) <= min( 1, Sum_i:[Pi] )

 which I see as a truncated INclusion-EXclusion principle for probabilities!
 Alternatingly adding and subtracting sums of P's of higher order tuples
 would make the ineqality on the lhs to flip-flop between the <= and >= .
 I have combined both inequalities into a SuperBonferroni principle :
!!!
  Max( Max_i:[Pi] , Sum_i:[Pi] - SumSum_j<k:[ Pjk ] ) <= P( U_i:[x_i] ) <=
                                                      <= min( 1, Sum_i:[Pi])
  For  m intersecting events is P(And_i(x_i)) :

   Max(0, 1 - Sum_i:[1-Pi] )  = Max[0, P(~(U_i:[~x_i])) as-if independent] =
 = Max(0, Sum_i:[Pi] -m +1)) <= P(And_i(x_i)) <=  min( Pi )

  which can be proven via DeMorgan's rule. These inequalities tell us that
  - few low  probability events have still a  low-probability union,
  - few high probability events have still a high-probability intersection.

  These inequalities are vital for checking probabilities or proportions
  given, generated or computed. Empirical P's should satisfy 0 < P(.) < 1,
  unless there are clear "structural zeros or ones" present, otherwise
  they should be smoothed, which will make them better estimates anyway.

  The subset-based measures of implication aka entailment:
  0  <=     Px - Pxy    <= 1  measures how little the event  x implies y,
   hence 1/(Px - Pxy)         measures how much   the event  x implies y
                                 with maximum = oo for Pxy = Px ;
  0  <= Pxy/Px = P(y|x) <= 1  measures how much   the event  x implies y
                                 with maximum = 1  for Pxy = Px ;
!!  recall that (x implies y) == (~y implies ~x)  in logic;  probabilistic
             is   1 - P(x,~y)  =  1 - (Px - Pxy)  { 1 - 0 iff Pxy=Px }

  P( x| y) = sensitivity (coined by Yerushalmy in 1947)
           = a/(a+c)  =  true positive ratio
  P(~x|~y) = specificity (coined by Yerushalmy in 1947)
           = d/(d+b)  =  true negative ratio
  P( x|~y) = 1 - specificity = false positive ratio = false alarm rate
           = b/(b+d)
  P( y| x) = a/(a+b) = positive predictivity (of x from y )
  P(~y|~x) = d/(d+c) = negative predictivity
  LR(x: y) = LR+ = positive likelihood ratio
                 =    P(x|y)/P(x|~y)   =      P(x|y)/( 1 -  P(~x|~y)   )
                 = (a/(a+c))/(b/(b+d)) = sensitivity/( 1 - specificity )

             LR- = negative likelihood ratio
                 =   P(~x|y)/P(~x|~y)  =
                 = (c/(c+a))/(d/(d+b)) = ( 1 - sensitivity )/specificity

  RR(y:x) = relative risk  =  risk ratio
    =   P(y|x) /  P(y|~x)
    = (Pxy/Px) / ((Py - Pxy)/(1 - Px))       which I rearranged into :
!!! =  Pxy*[1/(Py - Pxy)]*(1 - Px)/Px
!!! = [Pxy   / P(~x,y)  ]*(1 - Px)/Px        which I interpret as follows:
  RR(y:x)
  + INCreases with Pxy in general (due to numerator and denominator);
  + INCreases with Pxy approaching Py in particular, eg when (y implies x)
        fully then RR(y:x) = oo ie infinity;
! - DECreases with INCreasing Px in (1 - Px)/Px  which meaningfully measures
    our LOW/high surprise when a COMMON/rare event x occurs (more below).
    Note that P(x|y) = Pxy/Py as a measure of how much is the occurrence of
                y Sufficient for x, hence  of how much is the occurrence of
                x Necessary  for y, is not an explicit function of Px, hence
!!            P(x|y) cannot discount the lack of surprise like RR(y:x) does.
!!!
  RR(y:x) = CoOccur(x,y) * (y implies x) * SurpriseBy(x)
                                         { range [0..oo) for surprise too }
    or:   =              ( Pyx/P(y,~x) ) * SurpriseBy(x)
!!! ie:   =       LikelyThanNot(y: x)   * SurpriseBy(x)

         where LikelyThanNot(y:x) is visualized by my squashed Venn
         diagram (imagine two overlapping pizzas or pancakes x and y,
         and view them from aside) :

    xxxxxxxxxxxxx---    length of xx..x  = Px
          yyyyyyyyyy    length of yy..y  = Py
                        length of ---    = P(~x,y) = Py - Pxy
    It is meaningful to contrast the overlap  Pxy  against the underlap
        P(~x,y),  eg to contrast them relatively as their ratio :
    Pxy/P(~x,y)  >=<  1       where >=< stands for >, =, <, >=, <=, <>
    ie   P(x,y)  >=<  P(~x,y)
    ie   P(x|y)  >=<  P(~x|y) , so that eg for the > we say that :
    x occurs More Likely Than Not   if y occurred, or we say equivalently :
    x occurs More Likely Than Not with y , which both capture our thinking.

  + DE/INcreases with IN/DEcreasing Px; this is meaningful, because our
!!!    surprise value of x DIScounts the "triviality effect" of Px =. 1 :
!!   if Px =. 1 then  Pxy = Py too easily occurs,  and RR(y:x) = 1/0 = oo.
!!   If Py =  1 then  Pxy = Px and P(y,~x)=P(~x) hence RR(y:x) = 1/1 = 1,
       indeed, if all are ill, there can be no risk of becoming ill.
       Surprise value of x DE/INcreases with IN/DEcreasing Px in general;
       (1 - Px), 1/Px, hence also my (1 - Px)/Px measures our surprise by x.
!!     My new measure  P(x|y)*(1 - Px) = (y implies x)*(linearSurprise by x)
                                       =        [0..1]*[0..1] = [0..1]
       is simpler, but carries less meanings than RR(y:x).

! + is DOMINAted by the factor  1/(Py - Pxy)  for a given exposure Px ;
       this factor measures how much (y implies x). From this and from
         Pxy <= min(Px, Py), but not from "SurpriseBy", follows :
!!!    if Py < Px then RR(y:x) >= RR(x:y) ie LR(x:y),
!!!    if Py > Px then RR(y:x) <= RR(x:y) ie LR(x:y), where
                            the = may occur for x,y independent ie RR(:)=1,
       or if Pxy=0=RR(:), as my program Acaus3 asserts.
       That "SurpriseBy" is not decisive wrt RR(y:x) >=< RR(x:y), follows
       from the comparison of:
       (y implies x) = 1/(Py - Pxy)  vs  (1 - Px)/Px = SurpriseBy(x)
              ie (1 - 0)/(Py - Pxy)  vs  (1 - Px)/(Px - 0).

       Lets write Px = k*Py  to reduce RR(:) to just 2 variables Pxy, Py,
       and lets compare      RR(y:x) with RR(x:y) ie LR(x:y) :

       (Pxy/(Py - Pxy))*(1 - Px)/Px  >=<  (Pxy/(Px - Pxy))*(1 - Py)/Py
    ie:
       (k*Py - Pxy)/(Py - Pxy)  >=<  k*(1 - Py)/(1 -k*Py) = Dy in shorthand

!!     Pxy  >=<  Py*(Dy - k)/(Dy - 1) =
                 Py*[(1 - Py)/(1 - k*Py) - 1]/[(1 - Py)/(1 - k*Py) - 1/k]
    Checked:
       Solving for k the RR(y:x) = RR(x:y), where Px = k*Py , yields a
       quadratic equation with two distinct real roots k1, k2 :
       k1 = 1            ie     Px = Py   which obviously is correct
       k2 = Pxy/(Py*Py)  ie  Py*Px = Pxy  which holds for independent x,y .

  + is oo ie infinite for Py - Pxy = 0  ie  P(~x,y) = 0  ie  Pxy = Py
          in which case y implies x fully, because then y is a SubSet of x,
            ie whenever y occurs, x occurs too ; draw a Venn diagram.

  + is ASYMMETRICAL ie directed ie oriented wrt the events x, y (this unlike
        correlation coefficients and other symmetrical association measures)

  . is a relative measure, a ratio (while differences are absolute measures
                    which may mislead us since eg 0.93 - 0.92 = 0.03 - 0.02)

  - is a combined measure which inseparably MIXes measuring of two key
       properties:
     -    stochastic  dependence, which is a   symmetrical property, and
     - probabilistic implication, which is an Asymmetrical property, which
     both I see as necessary conditions for a possible CAUSAL relationship
     between x and y.   Hence RR(:) INDICATES potential CAUSAL TENDENCY;

  + has range [0..1..oo) with 3 opeRationally interpretable fixed points:
    RR(y:x)
      =  oo iff Pxy = Py  ie  (y implies x), ie possibly (x causes y) ;
      =   1 iff y and x are fully independent ie iff Pxy = Px*Py
      =   0 iff (Pxy = 0) and (0 < Px < 1)       ie disjoint events x,y
      ie  RR(:) =   0   means disjoint ie mutually exclusive events x,y
      0 < RR(:) <   1   means negative dependence or correlation of x,y
      1 < RR(:) <= oo   means positive dependence or correlation of x,y
!!    ie  RR(:) has a huge unbounded range for positively dependent x,y  vs
          RR(:) has a small  bounded range for negatively dependent x,y ,
          hence both subranges are not comparable; the positive subrange is
!!        much more SENSITIVE than the negative subrange. In this respect
!!        F(:) is BALANCED but has no simple interpretation of risk ratio.
      = 0/0 if  (Py = 0 or Py = 0 hence Pxy = 0 too).
      = 1   if  (Py = 1 hence Pxy = Px, P(x|y) = Px/1 ie independent x,y)
            then RR(y:x) = (1-Px)/(1-Px) = 1 ie independence.
!!    = 1   if  (Px = 1 hence Pxy = Py, P(y|x) = Py/1 ie independent x,y)
            then RR(y:x) = Pxy/(0/0) = Py/(0/0) numerically, which may
!!    seem to be undetermined, but as just shown, Px = 1 means that P(y|x)
      does not depend on Px, ie that  x,y  are independent (find 0/0 ).

   RR(y:x) = P(y|x)/P(y|~x)  where in many (not all) medical applications
       y is a health disorder, and x is a symptom.  But both { Lusted 1968 }
       and { Bailey 1965, p.109, quoted: } noted that :
      "P(y|x) will vary with circumstances (social, time, location), however
!!     P(x|y) will have often a constant value because symptoms are a
              function of a disease processes themselves, and therefore
       relatively INdependent of other external circumstances. ...
       so we could collect P(x|y) on a national scale, and collect Py on
       a [ local/individual ] space-time scale.".  The [loc/indiv] is mine.
       Therefore we should compute RR(y:x) indirectly via Bayes rule ie via
       P(y|x) = Py*P(x|y)/Px  where P(x|y) is "global" and more stable.

  + RR(y:x) has an important advantage over its co-monotonic but nonlinear
    transform  F(y:x). The simple proportionality of RR(y:x) can be used to
    (dis)prove confounding. Good explanations of confounding are rare, the
    best introduction is in { Schield 1999 } where on p.3 we shall recognize
    Cornfield's condition
                 P(c|a)/P(c|~a) > P(e|a)/P(e|~a)  as  RR(c:a) > RR(e:a)
    and Fisher's
                 P(a|c)/P(a|~c) > P(a|e)/P(a|~e)  as  RR(a:c) > RR(a:e).

    Be reminded that "contrary to the prevailing pattern of judgment",
    as { Tversky & Kahneman 1982, p.123 } point out, it holds, in my more
    general formulation :
              (  P(y|x) >=< P(y|~x) ) == (  P(x|y) >=< P(x|~y) ).
    Hence also
              ( RR(y:x) >=< 1       ) == ( RR(x:y) >=< 1       ),

    where  >=<  stands for a consistently used >, =, <, >=, <=, <> .

  + For 3 more properties see { Schield, 2002, p.4, Conclusions }.


  More on probabilities :

  Keep in mind that it always holds:
  P(.)      + P(~.)        = 1  eg  P(y|x) + P(~y|x) = 1 ; hence also:
  P(x or y) + P(~(x or y)) = 1  from which via DeMorgan's rule follows:
  P(x or y) + P(~x,~y)     = 1
  P(x or y) + Pxy  =     Px + Py  see the overlap of 2 Pxy in a Venn diagram
  hence   P(~x,~y) = 1 -(Px + Py - Pxy)

  P(~x,~y) =   P(~(x or y))  by DeMorgan's rule; he died in 1871; "his" rule
                             has been clearly described by Ockham aka Occam
                             aka Dr. Invincibilis in Summa Logicae in 1323 !
           = 1 - P(x or y) = 1 - (Px + Py - Pxy)  and, surprise :

! Pxy - Px*Py = Pxy*P(~x,~y) - P(x,~y)*P(~x,y)    from 2x2 table's diagonals
                      { with / the rhs would be Odds ratio OR , find below }
              = Pxy*(1 -Px -Py +Pxy) - (Px -Pxy)*(Py -Pxy)
              = cov(x,y) = covariance of 2 "as-if random" events x,y , or
                            indicator events aka binary/Bernoulli events.
  from which follows for independent events only :
  iff   Pxy - Px*Py = 0  ie  cov(x,y) = 0  ie
  iff   Pxy = Px*Py  (this is equivalent to 16 other equalities)
! then  Pxy*P(~x,~y) = P(x,~y)*P(~x,y)  ie products on 2x2 table's diagonals
  are equal; this I call the 17th condition of independence (find 17 below),
  which is equivalent (==) to any of the other 4 + 3*(8/2) = 16
  mutually equivalent (==) conditions of independence, like eg:

     ( Pxy = Px*Py   ) == (   P(x|y) = Px      ) == ( P(y|x) = Py ) ==
  ( P(y|x) = P(y|~x) ) == ( P(~y|~x) = P(~y|x) ) ==
  ( P(x|y) = P(x|~y) ) == ( P(~x|~y) = P(~x|y) ) == etc


  Only for independent x, y it holds, via Occam-DeMorgan's rule :

  P(~(~x,~y)) = 1 - (1 -Px)*(1 -Py) = Px + Py - Px*Py = P(x or y) for indep.

  More of the mutually equivalent conditions of independence are obtained by
  changing x into ~x, and/or y into ~y, or vice versa. Any consistent mix
  of such changes will produce an equivalent condition of independence for
  events, negated or not, simply because AN EVENT IS AN EVENT IS AN EVENT
  (with apologies to Gertrude Stein who spoke similarly about a rose :-)

  Changing the = into < or > in any of the 17 conditions of independence
  will create corresponding and mutually equivalent conditions of dependence
  which obviously are necessary but far from sufficient conditions for a
  causal relation between 2 events x, y. For example :

      ( Pxy    > Px*Py   ) == ( P(y|x) > Py     ) == ( P(x|y) > Px )
!! == ( P(y|x) > P(y|~x) ) == ( P(x|y) > P(x|~y) ) == etc.

  From all these 17 inequalities of the generic form lhs > rhs we can obtain
  some 6*17= 102 measures of DEPENDENCE simply by COMPARING or CONTRASTING :

  Da = lhs - rhs  are ABSOLUTE DEPENDENCE measures, eg P(e|h) - P(e|~h)
  Da is scaled [0..1]     for lhs > rhs , or  [-1..1]     in general,
           with 0 iff x,y are fully independent

  Dr = lhs / rhs  are RELATIVE DEPENDENCE measures, eg P(e|h) / P(e|~h)
  Dr is scaled [0..1..oo) with 1 iff x,y are fully independent.

  Rescalings : log(lhs / rhs) is scaled (-oo..0..oo)   in general;
     (lhs - rhs )/(lhs + rhs) is scaled  [-1..0..1], I call it kemenization,
   = (lhs/rhs -1)/(lhs/rhs +1);
   and        lhs/(lhs + rhs) is scaled  [0..1/2..1].

  Odds(.) =    P(.)/(1 - P(.))     =  1/( 1/P(.)     - 1 )
     P(.) = Odds(.)/(1 + Odds(.))  =  1/( 1/Odds(.)  + 1 )
  P(x| y)/P(~x| y) = P(x| y)/(1 - P(x| y))  = Odds(x| y)
  P(x|~y)/P(~x|~y) = P(x|~y)/(1 - P(x|~y))  = Odds(x|~y)
  P(x| y)/P( x|~y) = B(x: y)  =  LR(x: y)   = LR+ is a likelihood ratio
              where  B(x: y) is a simple Bayes factor = RR(x:y)

  Bayes rule in odds-likelihood form :

  Posterior odds on x if y   =   Prior odds  *  Likelihood ratio
    = Odds(x|y)              =     Odds(x)   *  LR(y:x)
    =    P(x|y)/   P(~x|y)   =  ( Px/P(~x) ) * ( P(y|x)/P(y|~x) )
    =    P(x|y)/(1 -P(x|y))  =    Px/(1 -Px) * ( P(y|x)/P(y|~x) )
    =         1/(1/P(x|y) - 1)


  In our 2x2 contingency table we have Odds ratio OR :

  OR       = Pxy*P(~x,~y) / [ P(x,~y)*P(~x,y) ] = (a/b)/(c/d) = a*d/(b*c)

  SeLn(OR) = sqrt( 1/a + 1/b + 1/c + 1/d ) = standard error of odds ratio OR

  cov(x,y) = Pxy*P(~x,~y) - [ P(x,~y)*P(~x,y) ]
           = Pxy - Px*Py
  but  OR <> Pxy /(Px*Py) , except when Pxy = Px*Py , or Pxy=0 .

  Relative risks RR(:) for the following 2x2 contingency table:

        |  e  | ~e  |         e = effect present;  ~e = effect absent
    ----|-----|-----|------
      h |  a  |  b  |  a+b    h = hypothetical cause present  (eg tested+ )
     ~h |  c  |  d  |  c+d   ~h = eg unexposed to environment (eg tested- )
    ----|-----|-----|------
        | a+c | b+d |  n

  RR( e: h) =   P(e|h)/ P(e|~h) = (a/(a+b))/(c/(c+d)) = a*(c+d)/((a+b)*c)
            = (Peh/Ph)/((Pe-Peh)/(1-Ph))  from which we see that RR(e:h)
!                 = oo if Pe=Peh ie (a+c)=a  ie  P(e,~h)=0 ie c=0

  Now recall that P(e|~h) + P(~e|~h) = 1, and get:
```

```
  RR(~e:~h) =   P(~e|~h)/P(~e|h)                      = (d/(c+d))/(b/(a+b))
            =   (1 - P(e|~h))/(1 - P(e|h))
            =   (1 - c/(c+d))/(1 - a/(a+b))           =  d*(a+b) /(b*(c+d))
!                = oo if Peh=Ph ie a=(a+b)  ie  P(h,~e)=0 ie b=0

  RR( h: e) =   P(h|e)/ P(h|~e) = (a/(a+c))/(b/(b+d)) = a*(b+d)/((a+c)*b)
            = (Peh/Pe)/((Ph-Peh)/(1-Pe))  from which we see that RR(h:e)
!                = oo if Ph=Peh ie (a+b)=a  ie  P(h,~e)=0 ie b=0

  RR(~h:~e) =   P(~h|~e)/P(~h|e)                      = (d/(b+d))/(c/(a+c))
            =   (1 - P(h|~e))/(1 - P(h|e))
            =   (1 - b/(b+d))/(1 - a/(a+c))           =  d*(a+c) /(c*(b+d))
!                = oo if Peh=Pe ie a=(a+c)  ie  P(e,~h)=0 ie c=0
  ie:
  for c=0 are RR( e: h) = oo = MAXImal = RR(~h:~e)
  for b=0 are RR( h: e) = oo = MAXImal = RR(~e:~h)
  for a=0  is RR( e: h) =  0 = minimal = RR( h: e)
  for d=0  is RR(~h:~e) =  0 = minimal = RR(~e:~h)

! RR(e:h)*RR(~e:~h) = RR(h:e)*RR(~h:~e) = Peh*P(~e,~h)/( P(e,~h)*P(h,~e) )
                                        = Peh*P(~e,~h)/( P(~e,h)*P(~h,e) )
```

which clearly are identical.
While these equations hold in general, you might like to meditate upon why the 17th (find above) condition of independent x, y consists from the same components.

```
  If you search www for "relative risk"          , you will get 160k hits;
  if you search www for "relative risk" RR       , you will get  28k hits;
  if you search www for "confidence interval" CI , you will search well.
```

-.-

+More tutorial notes on probabilistic logic, entropies and information :

Stan Ulam, the father of the H-device (Ed Teller was the mother) used to say that "Our fortress is our mathematics." I say that here "Our fortress is our logic." Elementary probability theory is strongly isomorphous with the set theory, which is strongly isomorphous with logic. There are 16 Boolean functions of 2 variables, of which 8 are commutative wrt both variables. For the purposes of inferencing we should use ORIENTED ie DIRECTED ie ASYMMETRIC functions only. From the remaining 8 asymmetric logical functions 4 functions are of 1 variable only, so that only 4 asymmetric functions remain for consideration : 2 implications and 2 inhibitions, which are pairwise mutually complementary. ASYMMETRY is
!!   easily obtained even from symmetrical measures of association (or dependence) by normalization with a function of one variable only, eg :

```
   (Pxy -Px*Py)/(Px*(1-Px))  is 0 iff x,y are independent
    = cov(x,y)/var(x)
    = beta(y:x)
    = slope of a probabilistic regression line Py = beta(y:x)*Px + alpha(y:x)
    =   (P(y|x) -  Py)/(1-Px)
    =    P(y|x) - P(y|~x)    is the numerator of F(y:x) below
    =  ARR(y:x) = absolute risk reduction (or increase if negative).
```

Many measures of information are easily obtained by taking expected value of either differences or ratios of the lhs and rhs taken from a dependence inequality lhs > rhs mentioned above. For example we could create :

```
   SumSum[ Pxy * Dr(y:x) ]  where  Dr  is a relative dependence measure like
                            eg RR(y:x), but a single Dr = oo would make the
```

```
                         whole  SumSum = oo, hence it is better to use
   SumSum[ Pxy * Da(y:x) ]  where  Da  is an absolute dependence measure, eg

   SumSum[ Pxy *( P(y|x) - P(y|~x) ) ],  or  SumSum[ Pxy * F(y:x) ] .

   Knowing that ( P(y|x) - P(y|~x) ) = beta(y:x) = dPy/dPx,  and
   knowing that Integral[ dx*( dPx/Px)^2 ] = Fisher's information, I did
   realize that my :
!!                  SumSum[ Pxy * (F(y:x))^2 ]  could serve as an

   quasi-Fisher-informatized RR [ find "my 1st interpretation" of F(y:x) ].

   A particularly nice & meaningfully asymmetrical (wrt variables X, Y)
   information is my favorite :

  �Cont(X;Y) = Cont(X) - Cont(X|Y)  ==  Gini(X;Y) = Gini(X) - Gini(X|Y)
    = Sum[ Px*(1 - Px) ] -   SumSum[  Pxy*(1-P(x|y)) ]
    = 1 -  Sum[ (Px)^2 ] - ( 1 - SumSum[ Pxy*P(x|y)  ] ) hence :
!!  = SumSum[ Pxy*( P(x|y) - Px ) ]          my semantically clearest form
!!  =     Expected[ P(x|y) - Px   ]          ie average dependence measured
      by abs. difference P(x|y) - Px , which is asymmetrical wrt x, y, and
      is but 1 of the 2*17 = 34 possible simple measures of association;

    = SumSum[ (square(Pxy - Px*Py)) / Py ] ,  compare it with Phi^2

   1-Cont(X) = Sum[ Px*Px] = E[Px] = expected probability of variable X
             =     expected probability of success in guessing events x
             =   long-run proportion of correct predictions of events x
             = concentration index by Gini/Herfindahl/Simpson (S. was a WWII
        codebreaker like I.J.Good and Michie; they called it a "repeat rate")

     Cont(X) = 1 - Sum[(Px)^2]  =  expected improbability of variable X
             =   expected error or failure rate eg in guessing events x

   0 <= Cont(;) and 1 - Cont(;)  <= 1  ie they saturate like P(error), while
                                 Shannon's entropies have no upper bound.
   Btw, log(.) fits with the physiological Weber-Fechner law. E.g. sound is
   measured on a log-scale in decibels, and so is the pH-factor (0..7..14 =
   max. alkalic). Logs work even in psychenomics, as you will feel less
   than twice as happy after your salary or profits were doubled :-) For more
   on infotheory in physiology see the nice book { Norwich 1993 }.

   Cont(;) has been called many names, eg quadratic entropy or parabolic
   entropy. Cont(;) gives provably better, sharper results than Shannon's
   entropy for tasks like eg pattern classification in general, and
   diagnosing, identification, prediction, forecasting, and discovery of
!  causality in particular. These tasks are naturally ASYMMETRICAL requiring
   Cont(X;Y) <> Cont(Y;X),  while Shannon's mutual information
      I(X:Y)  =    I(Y:X) = SumSum[ Pxy*( log(Pxy/(Px*Py)) ]
   is clearly symmetrical wrt the variables X, Y.

   Cont(X:Y)/Cont(X) = TauB in { Goodman & Kruskal, Part 1, 1954, p.759-760 }
   where they semantized their TauB as "relative decrease in the proportion
 ! of incorrect predictions".  See my Hint 2 & Hint 7 on WWW.MATHEORY.COM .
   { Agresti 1990, p.75 } tells us that for 2x2 contingency tables Kruskal's
   TauB equals Phi^2 :

   X^2 = mean square contingency { Kendall & Stuart, chap.33, p.555-557 }
       = n * SumSum[ square(Pxy - Px*Py)/(Px*Py) ] is my probabilistic form

   Pearson's contingency coefficient = sqrt[ X^2 / ( n + X^2 ) ].
   Phi^2  = (X^2)/n               compare it with the last lorm of Cont(X;Y)
          = SumSum[(square(Pxy - Px*Py))/(Px*Py) ]  in my probabilistic form
          = SumSum[ Pxy *   Pxy/(Px*Py)] - 1        is a symmetrical expected
                                    value like Shannon's mutual information :
   I(X:Y) = SumSum[ Pxy*log(Pxy/(Px*Py)) ] = I(Y:X) in general; in particular
          = -0.5*ln(1 - corr(X,Y)) iff X, Y are continuous Gaussian variables
```

[(1 - Cont(X)]/Pz = E[Px]/Pz = surprise index for the event z within the variable X, as defined in { Weaver, Science and Imagination }. In 1949 Weaver co-authored Shannon's The Mathematical Theory of Communication.

Cont(.) was intended to measure "semantic information content" SIC. The key idea is that the LOWER the probability of an event, the MORE possibilities it ELIMINATES, EXCLUDES, FORBIDS, hence MORE its occurrence SURPRISEs us. { Kemeny 1953, p.297 } refers this insight to { Popper 1972, pp.270, 399, 400, 402 mention P(~x) = 1-Px as (semantic information) content SIC }, { Bar-Hillel 1964, p.232 } quotes "Omnis determinatio est negatio", ie Determinatedness is negation, ie "Bestimmen ist verneinen" by Baruch Spinoza (1632-1677), in 1656 excommunicated from the synagogue in Amsterdam. Btw, Occam was excommunicated from the Church in 1328 :-). Stressing elimination of hypotheses or theories is Popperian refutationalism.

In principle any decreasing function of Px will do, but (1-Px) is surely the simplest one possible, simpler than Shannon's log(1/Px) = -log(Px). By combining (1 - Px) with 1/Px, I constructed SurpriseBy(x) = (1 - Px)/Px only to find it implicit or hidden inside RR(y:x), after my rearrangement of atomic factors in RR(:). Note that Sum[Px*1/Px] would not work :-)

For more on Cont(;) see { Kahre, 2002 } in general, and my (Re)search hints Hint2 & Hint7 there on pp.501-502 in particular (also on www.matheory.info). I could write a book(let) on Cont(.) but have to cut it here.

.-

Boolean logic is strongly isomorphous with the set theory wherein X implies Y whenever X is a subset of Y. Since the probability theory is also strongly isomorphous with the set theory, we see that for the < as the symbol for both "a subset of" and for the "less than", it is obvious that since [P(x|y) > P(y|x)] == [Py < Px] and v.v. ,

```
!!! Py < Px makes (y implies x) ie (x necessary for y)  more plausible, while
!!! Px < Py makes (x implies y) ie (y necessary for x)  more plausible,
```

which can be easily visualized with a Venn (aka pancakes or pizza) diagram.

```
    0 <= P(y|x)  = Pxy/Px    <= 1  measures how much   the event  x implies y
                                              with maximum = 1 for Pxy = Px ;
    0 <= P(x,~y) = Px - Pxy  <= 1  measures how little the event  x implies y
 or: 1 - P(y|x) = (Px - Pxy)/Px    measures how little the event  x implies y
 or:            1/(Px - Pxy)       measures how much   the event  x implies y
```

Also recall that the Bayesian probabibility of a j-th hypothesis x_j , given a vector of cue events y_c ie y..y, is (under the assumption of independence) computed by the Bayes chain rule formula based on the product of P(y|x) , the higher the more probable the hypothesis x_j :

```
   P(x_j, y..y) =. P(x_j)*Product_c:( P(y_c | x_j ) ;  dont swap y, x !
```

A cue event = a feature/attribute/symptom/evidential/test event
" x implies y " in plaintalk :

```
   " If x then y " is a deterministic rule, which in plain English says that
   " x always leads to y " ie  Px - Pxy = 0   ie   Px = P(x, y);  or:
   " It is not so that (x and not y) occur jointly" ie  P(x,~y) = 0 ;
                 note that Pxy + P(x,~y) = Px , hence   P(x,~y) = 0 and
                      the  Pxy = P(x) are equivalent indeed;
   which is the deterministic (extremely perfect or ideal) case
   which literally translates into the probabilistic formalisms:

    (x implies  y)  ==  ~(x,~y)  in logic, ie  1 - P(x,~y) = 1 - (Px - Pxy)

                      or, the smaller the P(x,~y), the more  x implies y ,
    hence another measure of probabilistic causal tendency (y causes x) :

    conv( x --> y)  = Px*P(~y)/P(x,~y) = (Px - Px*Py)/(Px - Pxy)
                    =       Px/P(x|~y) =  P(~y)/P(~y|x)
        is 1 iff x,y independent; and where:
      + the larger the Pxy <= min(Px, Py), the more the (x implies y),    and
      + the closer the Pxy  is to  Px*Py , the more independent are x, y, and
        the closer the conv(:) to 1 which is the fixed point for independence

        ( y -->  x) == (~x --> ~y) where --> is "implies" in logic; here too:

    conv( y -->  x) =  Py*P(~x)/P(y,~x) = (Py - Px*Py)/(Py - Pxy) =
  = conv(~x --> ~y) =  P(~x)*Py/P(~x,y)
                                 so their equality is logically ok, but it is
!!! UNDESIRABLE FOR A MEASURE OF CAUSAL TENDENCY. Q: why?  A: because eg:
   "the rain causes us to wear raincoat" is ok, but "not wearing a raincoat
    causes no rain" makes NO SENSE as the Nobel prize winner Herbert Simon
    pointed out in { Simon 1957, p.50-51 }. This undesirable equality does
    not hold for LR(:), RR(:) and its co-monotonous transformations like eg
    W(:) and F(:).

  (x inhibits y)  ==  (~x, y) == ~(~(~x, y)) == ~(y implies x) in logic,
                           is equivalent to  "y does not imply x" ;
                  == P(~x, y) =  Py - Pxy =   x inhibits y  (probabilistic)
                  ==    in plaintalk "Lack of x leads to y ",
                                   because in the perfect case we get
            ideally  P(~x,~y) = 0  is the deterministic, extreme case;
          note that  P(~x,~y) = 0  is not equivalent to Pxy = Py, because
        by DeMorgan  P(~x,~y) = 1 - (Px + Py - Pxy) = P(~(x or y)) always,
              hence  P(~x,~y) = 0 ie Px + Py - Pxy = 1 ie Px + Py = 1 - Pxy

             Recall  P(~x,~y) + P(~x, y) = P(~x)   always
                     P(~x,~y) + P( x,~y) = P(~y)   always

  inh0(x:y) =  P(~x,y)/(P(~x)*Py) = (Py -Pxy)/(Py -Px*Py) scaled [0..1..oo)
            = 0  iff Pxy = Py
            = 1  iff Pxy = Px*Py  ie iff x,y independent
            = oo iff Px  = 1

  inh1(x:y) = (  P(~x,y) - ( P(~x)*Py)) / (    P(~x,y) + ( P(~x)*Py) )
            = ((Py -Pxy) - (Py -Px*Py)) / ( ( Py -Pxy) + (Py -Px*Py) )
            = (   Px*Py  -  Pxy       ) / ( 2*Py -Pxy        -Px*Py  )
            = (   Pxy    -  Px*Py     ) / (       Pxy + Px*Py -2*Py  )
            = [-1..0..1]         by my kemenyzation

  inh1(y:x) = ( Px*Py - Pxy ) / ( 2*Px -Pxy -Px*Py )
            = ( Pxy   - Px*Py)/ ( Pxy +Px*Py -2*Px )

  x implies y == ~(y inhibits x), hence it should hold:

 -inh1(y:x) = (x implies y) = caus1(x:y), and indeed, it does hold
```

```
            = ( Pxy - Px*Py ) / ( 2*Px -Pxy -Px*Py ), see caus1(:) below.

 Consider again: "Lack of x (almost) always leads to y ".  Clearly, it
   would be wrong to tell somebody with x and y that x caused y .  Hence
   Pxy  alone cannot measure how much the x causes y, but P(y|x) could.
   Alas, P(y|x) = Pxy/Px is not a function of Py, and we believe that it is
   wise to have measures which are functions of all 3  Pxy, Px and Py :

    conv(x --> y) =  Px*P(~y)/P(x,~y)         the larger the more causation,
                                       due to small P(x,~y) co-occurence
                  = (Px - Px*Py)/(Px - Pxy)    is 1 if x, y are independent;
                  = [0..1..oo) , infinity oo iff Pxy = Px,  0 iff independ.

   conv2(x --> y) =  P(x implies y )/(  ~( Px*P(~y)) )  larger implies more
                  =       P(~(x,~y))/(  ~( Px*P(~y)) )
                  = ( 1 - P(  x,~y))/( 1 - Px*P(~y)  )  is 1 if independent;
                  = [1/2..1..4/3] , 1 iff x,y independent; 4/3 iff x imp y.

       From the   P(~x,~y) + P(~x, y) = P(~x)
                  P(~x,~y) + P( x,~y) = P(~y)
    for the case  P(~x,~y) = 0   holds  P(~y) = P( x,~y)  in which case
     conv(x --> y) =  Px                           <= 1 = independence,
    conv2(x --> y) = ( 1-P(~y) )/( 1 -P(~y)*Px )   <= 1 = independence;
    <= 1  is due to *Px , which always is  0 <= Px <= 1 .
    <= 1  in this case is good, because P(~x,~y) = 0  was shown to be
 !!!      equivalent to the (x inhibits y), hence x cannot imply y ,
          not even a little bit, ie causation must not exceed the point
          of no dependence ie point of independence, and indeed both
          conv(:) and conv2(:) are <= 1 in this case, which is good.

   An explanation and justification of the conv(:) measures:
    +  conv(:) = fun( Px, Py, Pxy )  ie fun of all 3 defining probabilities.
    +  conv has a fixed value if x, y independent , and also
         has a fixed value if x implies y 100% ,
         hence conv(:) has a decent opeRational interpretation.
    +  conv(x --> y) = extreme when  x implies y  100%
 !!!                        ie when Pxy = Px regardless of Py (draw a Venn)
           (x implies y) =     ~(x,~y)  in logic
                         = 1 - P(x,~y)  in probability
       { Brin 1997 } got rid of the outer negation ~ by taking the reciprocal
           value. On one hand this trick is not as clean as
 !!!       conv2(:), but on the other hand this trick makes the
 !!!       100% implication value an extreme value REGARDLESS of Py :

   conv(x --> y) =  Px*P(~y)/P(x,~y)  ,  the larger the more (x implies y)
                 =        Px/P(x|~y)  ,  is 1 iff independent x,y
     =  Px*(1 - Py)/(Px - Pxy)
     = (Px - Px*Py)/(Px - Pxy)  is  1  if Pxy = Px*Py  ie 100% independence,
                                is  oo if Pxy = Px     ie 100% (x implies y)
                                    oo  needs a precheck for an overflow;
                  numerically is 0/0 if  Py = 1 ie Pxy = Px  (overflow) but:
            correct logically is  1  if  Py = 1 as Pxy = Px*Py ie x,y indep.

   Or its reciprocal (since Pxy = Px is possible, while Py < 1) :
       (Px - Pxy)/(Px - Px*Py)        ,  the smaller the more (x implies y) :
                                is  1  if Pxy = Px*Py  ie 100% independence,
                                is  0  if Pxy = Px     ie 100% (x implies y)
                  numerically is 0/0 if  Py = 1 ie Pxy = Px  (overflow) but:
            correct logically is  1  if  Py = 1 as Pxy = Px*Py ie x,y indep.

   Conv(:) kemenyzed by me to the scale [-1..0..1] becomes
```

```
  conv1(x --> y)  = ( Px*P(~y) - P(x,~y) ) / ( Px*P(~y) + P(x,~y) )
                  = ( Px       - P(x|~y) ) / ( Px       + P(x|~y) )
           = ( Pxy - Px*Py)/(2*Px - Pxy - Px*Py)    =  -inh1(y:x)  above;
           =           ( P(~y) - P(~y|x) ) / ( P(~y)    + P(~y|x) )

     which kemenyzed to the scale [0..1/2..1] becomes :

  conv3(x --> y)  =   Px*P(~y) / ( Px*P(~y) + P(x,~y) )


  or based on counterfactual reasoning (cofa) IF ~x THEN ~y :

  cofa1(~x --> ~y) = ( P(~x)*Py  - P(~x, y) ) / ( P(~x)*Py  + P(~x, y) )
                   = ( Py -Px*Py - Py +Pxy  ) / ( Py -Px*Py + Py -Pxy  )
             = ( Pxy - Px*Py )/( 2*Py -Pxy -Px*Py)  =  -inh1(x:y)  above,

             ie  not( x inhibits y ).

  cofa0(~x --> ~y) =   P(~x)*Py / P(~x, y)
                   =  (Py -Px*Py)/(Py -Pxy)  =  conv(y --> x)  above,

    and indeed, in logic (~x <== ~y) == (y <== x);  the <== means "implies"
   (and also it means "less then" if applied to 0 = false, 1 = true)

  F(~x:~y) == F(~x <== ~y)
   = ( P(~x|~y) - P(~x|y) ) / ( P(~x|~y) + P(~x|y) ) = -F(~x:y)

    They all look reasonable, and all are scaled to [-1..0..1].
 Q: which one do you like, if any, and why (not) ?

  A mathematically more rigorous alternative to conv(x --> y) is my
  conv2(x --> y)  which does not suffer from the dangers of an overflow,
  employs the exact probabilistic (x implies y) =  1 - P(x,~y)  derived
     from the exact logical       (x implies y) ==     ~(x,~y).
  Since we wish to have a fixed value for the independence of events x, y,
  the exact implication form  1 - P(x,~y)  suggests to compare it with
  the negation of the fictive ie as-if independence term as follows:

  conv2(x --> y) =   P(x implies y)/( x,y independ  )   larger implies more
                 =       P(~(x,~y))/(  ~( Px*P(~y)) )
                 = ( 1 - P(  x,~y))/( 1 - Px*P(~y)  )   is 1 if independent
                 = ( 1 -(Px - Pxy))/( 1 - Px*(1-Py) )
                 = ( 1 - Px + Pxy )/( 1 - Px +Px*Py )   is 1 if Pxy = Px*Py

          This is  ( 1 - Px + Px  )/( 1 - Px +Px*Py )        if Pxy = Px
                               = 1/( 1 - Px*(1 -Py))   >= 1 if Pxy = Px,
          the larger the Px and the smaller the Py, the >> 1 is conv2(x;y).
                   When  Px = Pxy  (draw a Venn diagram)  ie if 100% implies
                   then the numerator is  1  ie maximal, but unlike in
!!                 conv(x --> y) , the denominator depends on Px and Py

-.-

+Rescalings important wrt risk ratio RR(:) :

    For positive u, v :     u/v   is scaled [  0..1..oo)         v <> 0
    and  W = ln(u/v)              is scaled (-oo..0..oo)         v <> 0
    and  F = (u - v  )/(u + v  )  is scaled [ -1..0..1 ], allows u=0 xor v=0
           = (1 - v/u)/(1 + v/u)  handy for graphing F=f(v/u)    u <> 0
           = (u/v - 1)/(u/v + 1)  handy for graphing F=f(u/v)    v <> 0
           = (u - v  )/(u + v  )  rescaling I call "kemenyzation" to honor
                                   the late John Kemeny, the Hungarian-American
                   co-father of BASIC, and former math-assistant to Einstein;

           = tanh(W/2) = tanh(0.5*ln(u/v))   due to { I.J. Good 1983, p.160
                                                 where sinh is his mistake }
    Since
          atanh( z ) = 0.5*ln( (1+z)/(1-z) )  for z <> 1,
    W = 2*atanh( F ) =     ln( (1+F)/(1-F) )  for F <> 1

    F0 = (F+1)/2 is linearly rescaled to [0..1/2..1],  1/2 for independence.

    W(y:x) = ln( P(y|x)/P(y|~x) )  is an information gain [see F(:) ]
           = ln( P(y|x) ) - ln(P(y|~x) ) is additive
           = ln( B(y:x) )               ie logarithmic Bayes factor
           = ln(RR(y:x) )   =    ln( relative risk of y if x )
         = ln(Odds(x|y)/Odds(x))
              is I.J. Good's "weight of evidence in favor of x provided by y".

    The advantage of oo-less scalings like [-1..0..1] or [0..1/2..1] is that
    they make comparisons of different formulas possible at all and more
    meaningful, though not perfect.  E.g. we may try to compare a value of
    F(:) with that of conv1(:) which is conv(:) kemenyzed by me.

    W(:)'s logarithmic scale allows addition (of otherwise multiplicable
         ratios) under the valid assumption of independence between y, z :

         W(x: y,z) =   W(x:y) + W(x:z)

     but when y,z are dependent we must use { I.J. Good 1989, p.56 } :

         W(x: y,z) =   W(x:y) + W(x: y|z)

    F(:)'s cannot be simply added, but can be combined (provided y, z are
         independent) according to { I.J. Good, 1989, p.56, eq.(7) } thus :

         F(x: y,z) = ( F(x:y) + F(x:z) )/( 1 + F(x:y)*F(x:z) )

     but when y,z are dependent we must use :

         F(x: y,z) = ( F(x:y) + F(x: z|y) )/( 1 + F(x:y)*F(x: z|y) )

    Seeing this, physicists, but not necessarily physicians, might recall
    that 2 relativistic speeds are combined into the resultant one by means
    of a regraduation function for relativistic addition of velocities  u, v
    into a single rapidity rap :

    rap = ( u + v )/( 1 + u*v/(c*c) )   where c is the speed of light.

    P(.|.)'s maximum = 1 corresponds to the unexceedable speed of light,
             in which case  rap  simplifies to our ( u + v )/( 1 + u*v ).

    This relativistic addition appears in:
    - { Lucas & Hodgson,  pp.5-13 } is the best on regraduation (no P(.)'s )
    - { Yizong Cheng & Kashyap 1989, p.628 eq.(20) }, good;
    - { Good I.J. 1989, p.56 }
    - { Grosof 1986, p.157 }    last line, no relativity mentioned;
    - { Heckerman 1986, p.180 } first line, no relativity mentioned.

-.-

+Correlation in a 2x2 contingency table is scaled to [-1..0..1] :

    corr(x,y) = [ a*d - b*c ]/ sqrt[ (a+b)*(a+c) * (b+d)*(c+d) ]
```

```
          = [ Pxy*P(~x,~y) - P(~x,y)*P(x,~y) ] / sqrt[ Py*Px*P(~x)*P(~y) ]

          = [ Pxy - Px*Py ] / sqrt[ Px*(1-Px) * Py*(1-Py) ]
          = cov(x,y) / sqrt( var(x) * var(y) )
          = correlation coefficient of binary ie Bernoulli ie indicator events
                x, y  is symmetrical wrt x, y

       r2 = square(corr(x,y))
          = ( cov(x,y)/var(x)) * ( cov(x,y)/var(y))
          =       beta( y:x ) * beta( x:y )                   -1 <= beta <= 1
          = (slope of y on x ) * (slope of x on y)
          = ( P(y|x) - P(y|~x) * ( P(x|y) - P(x|~y) )   for events x, y

          = coefficient of determination aka r^2 or r2
          = ( explained variance ) / ( explained var. + unexplained variance )
          = ( variance explained by regression ) / ( total variance )
          =        1 - ( variance unexplained ) / ( total variance )
       r2 is considered to be a more realistic (because less inflated) measure
          of correlation than the corr(.,.) itself (except for the sign).
       The key mean squared error equation from which the above follows is :

       MSE = variance explained + variance unexplained aka residual variance

       This MSE equation I call Pythagorean decomposition of the mean squared
       error MSE into its orthogonal partial variations.
       It is a sad fact that very few books on statistics and/or probability
       show the correlation coefficient between events.

        Yule's coefficient of colligation { Kendall & Stuart 1977, chap.33
               on Categorized data, p.539 } is also symmetrical wrt x, y:

        Y = ( 1 - sqrt(b*c/(a*d))   ) / ( 1 + sqrt(b*c/(a*d))   )
          = ( sqrt(a*d) - sqrt(b*c) ) / ( sqrt(a*d) + sqrt(b*c) )   kemenyzed
          = tanh( 0.25*ln( a*d/(b*c) ) )   my tanhyperbolization a la I.J. Good


       The formula for chi-squared (findable as X^2 , chisqr , chisquared ) :

       X^2 = Sum[ ( Observed - Expected^2 ) / Expected ]
           =    [ ( a - (a+b)(a+c)/n )^2 +
                  ( b - (a+b)(b+d)/n )^2 +
                  ( c - (a+c)(c+d)/n )^2 +
                  ( d - (b+d)(c+d)/n )^2  ]                  is exact,

       =.   n*(|ad - bc| -n/2)^2 /[ (a+b)(a+c)(b+d)(c+d) ]    Yates' correction
       =..  n*( ad - bc      )^2 /[ (a+b)(a+c)(b+d)(c+d) ]   may be good enough

    .-

      Although a meaningful interpretation of values is very important, it is
      equally important how it orders the values obtained from a data set, since
      we want a list of the pairs of events (x,y) sorted by the strength of their
      potential causal tendency :
      Note that :
        P(x|y) is the diagnostic predictivity of the hypothesis x from y effect
        P(y|x) is the     causal predictivity of the     effect y from x
        P(y|x) = Py*P(x|y)/Px = Pxy/Px   is the Bayes rule.

    The likelihood ratio aka Bayes factor in favor of the outcome (or hypothesis)
    x provided by the evidence (or predictor or cue or feature or effect) y,
    aka relative risk RR is :
```

```
 RR(y:x) == B(y:x)
         =  P(y|x) / P(y|~x)  =  (Pxy/Px)/( (Py - Pxy)/(1 - Px) )
         =  Pxy*(1 - Px) / (Px*Py - Px*Pxy)    caution! /0 if Pxy = Py :-(
         =      (1 - Px) / (Px*Py/Pxy - Px)
         = (Pxy - Px*Pxy)/( Px*Py - Px*Pxy )    which shows that:
                 B = 1  for Px*Py=Pxy  ie for independent x,y ;  and
                 B = oo for    Py=Pxy  ie "if y then x" ie  y implies x
         =  relative odds on the event x after the event y was observed
      !! = Odds(x|y)/Odds(x) =  posteriorOdds / priorOdds ;  { odds form }
         =  ( P(x|y)/(1-P(x|y)) )/( Px/(1-Px)     )
         =  ( P(x|y)*(1-Px)     )/( Px*(1-P(x|y)) )
            { note that (x|y) inverts into (y|x) via
                            Bayesian P(x)*P(y|x) = Pxy = P(x|y)*Py }
         =    P(y|x) / P(y|~x)                q.e.d.
    !!   =  ( Pxy/(Py - Pxy) )*( (1-Px)/Px )
    !!  shows that Py = Pxy  does mean that (y implies x) so that B(y:x) = oo
    !!   note that when (x causes  y)  then (y implies x) but not necessarily
    !!            vice versa; the y is an effect or outcome in general;
         = P(y|x)/( 1 - P(~y|~x) ) = B(y:x) because,
         = P(y|x)/P( y|~x)                  q.e.d.

  Lets compare RR(y:x) = P(y|x) / P(y|~x) = P(y|x) * (1-Px)/(Py - Pxy)
!!
     with  conv(y --> x) =   Py / P(y|~x) =  Py    * (1-Px)/(Py - Pxy)
                                          =  Py    *  P(~x)/P(y,~x)

  clearly RR(y:x) is more meaningful than the "conviction" by { Brin 1997 },
  though conviction is no nonsense either :
   + both RR(y:x) and conv(y:x) equal oo if Py=Pxy ie if y implies x
   + both RR(y:x) and conv(y:x) equal  1 if Pxy=Px*Py ie y, x are independent
   + both RR(y:x) and conv(y:x) equal  0 if Pxy=0  ie if y is disjoint with x
   +      RR(y:x) is relative risk, used within other meaningful formulas
   +      RR(y:x) <>   RR(~x:~y)  which is    good, while
   -    conv(y:x) == conv(~x:~y)  which is NO GOOD (find UNDESIRABLE above)

  B(~y:~x) =      P(~y|~x)  / P(~y|x)          ==  RR(~y:~x)
           = [1 - P( y|~x)] / [ 1 - P(y|x)] = [ P(~y,~x)/P(~y,x)]*Px/(1 - Px)
           = [ (1 -Py -Px +Pxy)/(Px -Pxy) ]* Px/(1-Px)
           = [ (1 -Py)/(Px-Pxy)  -1       ]* Px/(1-Px)
        !!        when  Px=Pxy  ie when (x implies y) then B(~y:~x) = oo
                  hence if we wish to use a B(~.:~.) instead of B( .: .),
               then we must swap the events x and y. For example instead of
  B( y: x) we might use :
  B(~x:~y) =      P(~x|~y)  / P(~x|y)          ==  RR(~x:~y)
           = [ (1 -Px)/(Py-Pxy)  -1 ]* Py/(1-Py)
        !!        when  Py=Pxy  ie when (y implies x) then B(~x:~y) = oo

  W(y:x) = the weight of evidence for x if y happens/occurs/observed
   =  ln( P(y|x)/P(y|~x) )   = Qnec(y:x)      { I.J. Good 1994, 1992 }
   =  ln( B(y:x) ) = logarithmic Bayes factor for x due to y
   =  ln(RR(y:x) )
  = ln(Odds(x|y)/Odds(x))

  W(~y:~x) = the weight of evidence against x if y absent { I.J. Good }
   =  ln( P(~y|~x)/P(~y|x) ) = Qsuf(y:x)       { I.J. Good 1994, 1992 }
   =  ln( B(~y:~x) )
   =  ln( (1 - P( y|~x)) / (1 - P( y|x)) )
   = -W(~y:x)

  W(:) = 2*atanh(F(:)) = ln((1+F)/(1-F))  for abs(F) <> 1
  B(a:b) = P(a|b)/P(a|~b) = (Pab/Pb)/((Pa-Pab)/(1-Pb))
         = oo iff Pab=Pa ie iff (a implies b)

  B(b:a) = P(b|a)/P(b|~a) = (Pab/Pa)/((Pb-Pab)/(1-Pa))
         = oo iff Pab=Pb ie iff (b implies a)

  Q:/Quiz: could comparing (eg subtracting or dividing) B(a:b) with B(b:a)
     show the DIRECTION of a possible causal tendency ??

  P(~b,~a) = 1 - (Pa + Pb - Pab) = P(~(a or b))  by DeMorgan's rule

  B(~b:~a) =  P(~b|~a)/P(~b|a) = (1 - P(b|~a))/(1 - P(b|a))
           = [P(~b,~a)/(1-Pa)] / [(Pa-Pab))/Pa]
           = oo iff Pab=Pa ie iff (a implies b) like for B(a:b) or W(a:b)

  which speaks against comparing ?(a:b) with ?(~b:~a) for the purpose of
  deciding the direction of possible causal tendency.


  C(y:x) = a measure of corroboration             { by Karl Popper }
   = (P(y|x) - Py     )/( P(y|x) + Py -   Pxy    )    { C-form 1   }
   = (   Pxy - Px*Py  )/( Pxy + Px*Py -   Pxy*Px )    { C-form 2   }
   = (P(y|x) - P(y|~x))/( P(y|x) + Py/P(~x) )      { compare w/ F-form 1 }
   = (cov(x,y)/var(x) )/( P(y|x) + Py/P(~x) )         { C-form 3a  }
   =          beta(y:x)/( P(y|x) + Py/P(~x) )         { C-form 3b  }

  F(y:x) == F(y <== x) = degree of factual support of x by y
   = primarily a measure of how much y implies x  { by John Kemeny }
   = (P(y|x) - P(y|~x))/( P(y|x) + P(y|~x)  )         { F-form 1   }
   = (   Pxy - Px*Py  )/( Pxy + Px*Py - 2*Pxy*Px )    { F-form 2   }
   = (cov(x,y)/var(x) )/( P(y|x) + P(y|~x)  )         { F-form 3a  }
   =          beta(y:x)/( P(y|x) + P(y|~x)  )         { F-form 3b  }
   =         tanh( 0.5*ln(P(y|x) / P(y|~x)) )         { F-form 4   }
   =         tanh( W(y:x)/2 )                         { by I.J.Good, fixed }
   = (difference/2) / average =    deviation/mean     { F-form 5   }
   = ( B(y:x) - 1 ) / ( B(y:x) + 1 )  is handy also for graphing F = fun(B)
   =  -F(y:~x)
   =  (Pxy/Px - (Py-Pxy)/(1-Px)) / (Pxy/Px + (Py-Pxy)/(1-Px)) hence:
   =  1     iff P(y|~x) = 0
         ie iff P(y,~x) = 0   ie iff Pxy = Py
         ie iff y implies  x   deterministically
         ie iff y leads to x   (always)
         ie  IF y THEN x       (always holds)
   =  0     iff x, y  are independent ;
   = -1     iff P(y|x)  = 0
         ie iff P(y,x)  = 0   ie iff x, y are mutually exclusive
where :
the F-form 1 is the original one by { Kemeny & Oppenheim 1952 },
 my F-form 2 is the de-conditioned one, and it does reveal that
             iff Pxy=Py (see the -2* ), ie iff y implies x, then F(x:y)=1.
 my F-form 3 reveals an important hidden meaning: beta(y:x) is the slope
    of the implicit probabilistic regression line
    of Py = beta(y:x)*Px + alpha(y:x) ;
the F-form 4 reveals that F(:) and Turing-Good's weight of evidence W(:)
             are changing co-monotonically
 my F-form 5 provides the most simple interpretation of F(:)

Unlike B(:) or W(:), the F(:) will not easily overflow due to /0.
The numerators tell us that for independent x, y it holds  C(:) = 0 = F(:).
C(:) stresses the near independence, while F(:), W(:), B(:) stress near
implication more than near independence. Try out an example with a near
independence and simultaneously with near implication.
```

```
  F(x:y) == F(x <== y) = degree of factual support of y by x
   = primarily a measure of how much x implies y
   = (P(x|y) - P(x|~y))/( P(x|y) + P(x|~y)  )                 { F-form 1 }
   = (   Pxy - Px*Py  )/( Pxy + Px*Py - 2*Py*Pxy )            { F-form 2 }
   = (Pxy/Py - (Px-Pxy)/(1-Py)) / (Pxy/Py + (Px-Pxy)/(1-Py))
  = -F(x:~y)

  note that iff Px=Pxy then F(x:y) = 1 == (x implies y) fully, and that
  this matches  Pxy/Px = 1 as maximal possible contribution to the product
  for P(x_j | y..y) computed by the simple Bayesian chain rule over y..y cues
  for P(x_j , y..y).  Clearly a product of Pxy/Px terms over a vector of cues
  y..y  may be viewed as a product of the simplest (x implies y) terms.

Rescaling F(:) from [-1..0..1] to [0..1/2..1] :

  F0(:) = ( F(:) + 1 )/2 , so that

  F0(x:y) = P(x|y)/( P(x|y) + P(x|~y) )
  F0(y:x) = P(y|x)/( P(y|x) + P(y|~x) )

Before we go further, we recall that F(:) is co-monotonical with B(:), and
    B(y:x) =  P(y|x) / P(y|~x)
           = (Pxy/Px)/( (Py - Pxy)/(1 - Px) )  {now consider INCreasing Pxy:}
     wherein  Pxy/Px     measures how much (x implies y) up to maximum = 1
     while  1/(Py - Pxy) measures how much (y implies x) up to maximum = oo
     hence iff Py = Pxy then B(y:x) = oo ie y implies x is measured by B(y:x)
          and:
   B(y:~x) = P(y|~x) / P(y|x)
           = ((Py - Pxy)/(1 - Px)) /(Pxy/Px)   {now consider DECreasing Pxy:}
      wherein  Py - Pxy  measures how much (y implies x) with maximum = Py
      while 1/(Pxy/Px)   measures how much (x implies y) with maximum = oo
                         for Pxy=0
     hence iff Pxy = 0 then B(y:~x) = oo

       F(y:x)
   =  (P(y|x) - P(y|~x))/(P(y|x) + P(y|~x))
   =  ( Pxy   - Px*Py  )/( Pxy   + Px*Py - 2*Px*Pxy )
   = -F(y:~x)

       F(y:~x)
   =  (P(y|~x) - P(y|x))/(P(y|~x) + P(y|x))
   =  (  Px*Py - Pxy   )/( Px*Py  + Pxy - 2*Px*Pxy )
   = -(  Pxy   - Px*Py )/( Pxy  + Px*Py - 2*Px*Pxy );
   = -F(y:x)

       F(~y:x)
   =  (P(~y|x) - P(~y|~x))/(P(~y|x) + P(~y|~x))
   =  (  Pxy   - Px*Py   )/(   Pxy  + Px*Py - 2*Px*(1 - Px + Pxy) )
   = -F(~y:~x)

       F(~y:~x)
   =  (P(~y|~x) - P(~y| x))/(P(~y|~x) + P(~y| x))
   =  (   Px*Py - Pxy     )/( Px*Py   + Pxy   - 2*Px*(1 - Px + Pxy) )
   = -(   Pxy   - Px*Py   )/(   Pxy   + Px*Py - 2*Px*(1 - Px + Pxy) )
   = -F(~y:x)

and the remaining 4 mirror images are easily obtained by swapping x and y.
Note that W(:) = 2*atanh(F(:)) = ln[ (1 + F(:))/(1 - F(:)) ] .

For example  F(~y:x) = -F(~y:~x)  would measure how much the hypothesis
   x  explains the unobserved fact y , like e.g. in common reasoning :
```

```
"if (s)he would have the health disorder x
    (s)he could NOT be able to do  y  (eg a body movement (s)he did)",
so that from a high enough  F(~y:x)  we could exclude the disorder  x  as
an unsupported hypothesis.

-.-

+Example 1xy:
  for  Px=0.1 , Pxy=0.1 , Py=0.2 ,  visualized by a squashed Venn diagram

        xxxxxxxxxx
        yyyyyyyyyyyyyyyyyyyy

   are P(x|y)=0.5 ie "50:50" ;  P(x|~y)=(0.1 -0.1)/(1 -0.2) = 0 ie minimum
       P(y|x)=1.0 ie maximum ;  P(y|~x)=(0.2 -0.1)/(1 -0.1) = 1/9
   and
       corr(x,y) = cov(x,y)/sqrt[ var(x) * var(y) ]
                 = (Pxy - Px*Py)/sqrt[ Px*(1-Px) * Py*(1-Py) ]
                 = 0.7 is the value of the correlation coefficient between
                          the events x, y
   caus1(x:y) = ( Pxy - Px*Py ) / ( 2*Px -Pxy -Px*Py ) = 1
       F(x:y) = (0.5 - 0)/(0.5 + 0) = 1
       B(x:y) =    P(x|y)/P(x|~y)   = 0.5/0 = oo = infinity
       clearly the rule IF x THEN y  cannot be doubted ;

   but what do we get when we swap the roles of x, y ie when our observer
   will view the situation from the opposite viewpoint ? This can be done
   by either swapping the values of Px with Py, or by computing F(y:x) :

+Example 1yx:
   is  F(y:x) = (1 -1/9)/(1 + 1/9) = 0.8  is too high for my taste
   caus1(y:x) = (Pxy - Px*Py)/(2*Py -Pxy -Px*Py) = 0.29 is more reasonable,
      as Pxy/Py = 0.5 hence y doesnt imply x much (although x implies y fully
      as Pxy/Px = 1 );
       B(y:x) = P(y|x)/P(y|~x) = 1/((0.2 - 0.1)/0.9) = 9 is (too) high.

!!! Conclusion: for measuring primarily an implication & secondarily
                dependence, B(y:x) and F(y:x) are not ideal measures.

!!! Note: if Px < Py then x is more plausible to imply y, than vice versa;
          if Py < Px then y is more plausible to imply x, than v.v.

+Example 3:    x  = drunken driver ;     y = accident
          P(y| x) = 0.01   = P( accident y caused by a drunken driver  x )
          P(y|~x) = 0.0001 = P( accident y caused by a sober   driver ~x )
      is how { Kahre 2002, p.186 } defines it; obviously P(y|x) > P(y|~x).
 Note that without knowing either Px or Py or Pxy, we cannot obtain the
 probabilities needed for   caus1(x:y), F(x:y) and F(~x:~y), ie we can
 compute B(y:x), F(y:x) and caus1(y:x) only :

  beta(y:x) = P(y|x) - P(y|~x) =.  0.1  is the regression slope of y on x ,
                                        is misleadingly low.
  B( y: x) =  P(y|x) / P(y|~x) = 100    ie (y implies x) very strongly.
  F( y: x) =   (B-1) / (B+1)   =   0.98  =. 1 = F's upper bound
  F( y: x)  measures how much an accident y implies drunkenness x
            (obviously an accident cannot cause drunkenness).

  B(~y:~x) = (1-P(y|~x))/(1-P(y|x)) = 1.01
  F(~y:~x) =       (B-1)/(B+1)      = 0.005 =. 0 = F's point of independence
  F(~y:~x)  measures how much an absence of an accident y implies that
            a driver is not drunk. Here is my CRITICISM of such formulas:
According to I.J. Good, "The evidence against x if y does not happen" can
```

```
also be considered as a possible measure of x causes y . It is based on
   COUNTERFACTUAL reasoning "if absent y then absent x", which I denote as
!! "Necessitistic" reasoning.  I am dissatisfied with the sad fact that his
   formulation leads to formulas which are not zero when Pxy = 0 ie when x, y
are DISjoint. If the above explained notion of Necessity is to be taken
seriously, and I think it should be, then Good's formulation is not good
enough.

  F(~y:~x) == F(~y <== ~x)
   =  measures how much ~y implies ~x  =
   =  (P(~y|~x) - P(~y| x))/(P(~y|~x) + P(~y| x))                        (1)
   =  (   Px*Py - Pxy     )/( Px*Py   + Pxy   - 2*Px*(1 -Px + Pxy) )
   = -(   Pxy   - Px*Py   )/(   Pxy   + Px*Py - 2*Px*(1 -Px + Pxy) )
   = (B(~y:~x) - 1)/(B(~y:~x) + 1)
   = ((1 - P(y|x~)) - (1 - P(y|x )))/((1 - P(y|x~)) + (1 - P(y|x)))  from (1)
   = (   - P(y|x~)  +      P(y|x ) )/( 2 - P(y|x~)       - P(y|x) )
   = (     P(y|x )  -      P(y|x~) )/( 2 - P(y|x~)       - P(y|x) )
   = -F(~y:x)

  F(~x:~y) == F(~x <== ~y)
   = ( P(~x|~y) - P(~x|y) )  / ( P(~x|~y) + P(~x|y) )  = -F(~x:y)

  Kemenyzed are:

  Knec( y: x) = ( P(y|x) - P(y|~x) ) / ( P(y|x) + P(y|~x) )
              = ( cov(x,y)/var(x)  ) / ( P(y|x) + P(y|~x) )

  Ksuf( y: x) = (  (1 - P(y|~x))  - (1 - P(y|x)) )
               /(  (1 - P(y|~x))  + (1 - P(y|x)) )
              = ( P(y|x) - P(y|~x) ) / ( 2 -(P(y|x) + P(y|~x)) )
              = ( cov(x,y)/var(x)  ) / ( 2 -(P(y|x) + P(y|~x)) )
-.-

+Folks' wisdom :

!!! Caution: causation works in the opposite direction wrt implication. This
             is so, because ideally an effect  y  implies a cause  x, ie
       a cause  x  is necessary for an effect  y. See the short +Introduction
again. In what follows here it may be necessary to swap  e, h  if we want
causation. Since different folks had different mindwaves I tried not to mess
with their formulations more than necessary for this comparison.

The notions of probabilistic Necessity and Sufficiency on the event level
have been quantified differently by various good folks' wisdoms.

E.g. from { Kahre 2002, Fig.3.1, Fig.13.4 + txt } follows :

 - if     X is a   subset of Z  ie  Z is a SuperSet of X  ie  all X are Z
     then X is Sufficient (but not necessary)  for Z      ie  X implies Z.
       ie Z is a consequence of X  ie IF X THEN Z rule holds, I say.

 - if     Y is a SuperSet of Z  ie  Z is a   subset of Y  ie  all Z are Y
     then Y is Necessary  (but not sufficient) for Z      ie  Z implies Y
       ie Y is a consequence of Z  ie IF Z THEN Y rule holds, I say;

For 2 numbers x, y it holds (x    <      y) == ( y      >      x) , or v.v.
For 2 sets    X, Y it holds (X subset of Y) == ( Y superset of X) , or v.v.
For 2 events  x, y it holds (x implies   y) == (Py      >     Px) , or v.v.

For 2 events we may like to answer the Q's (and from the above follow A's) :
Q: How much is x Sufficient for y ?   A: as much as y is Necessary  for x .
Q: How much is x Necessary  for y ?   A: as much as y is Sufficient for x .
```

```
Q: If Pxy=0 ie x, y are disjoint  ?   A: then Suf = 0 = Nec  must hold.

Lets use:  e = evidence, effect, outcome;  h = hypothesised cause (exposure)
Hence eg P(e|h) is a NAIVE, MOST SIMPLISTIC measure of Sufficiency of h for e
because  P(e|h) = 1 = max iff Peh = Ph  ie  Ph - Peh = 0 = P(~e,h)  ie
iff h is a subset of e, ie iff h implies e then is h Sufficient for e.
Note that (h Sufficient for e) == (h subset of e), and
         (h Necessary  for e) == (e subset of h), hence :
P(h|e) measures how much is h Necessary  for e, and
P(e|h) measures how much is h Sufficient for e.
Lets compare these with those now corrected in { Schield 2002, Appendix A } :
P(e|h) = S = "Sufficiency of exposure h for case e"
P(h|e) = N = "  Necessity of exposure h for case e" (find NAIVE , SIMPLISTIC )

!! Caution: the suffixes nec, suf in ?nec, ?suf as used by various authors
            say nothing about which event is necessary for which one, if
   the authors do not use ?(y:x) and do not specify what these parameters
   mean. I recommend the  ?(y:x) to mean that (y implies x) ie
   (y suffices for x) which is equivalent to  (x necessary for y).


Folk1: { Richard Duda, John Gaschnig & Peter Hart: Model design in the
        PROSPECTOR consultant system for mineral exploration,
   in { Michie 1979, pp.159 } , { Shinghal 1992, chap.10, pp.354-358 } and
   in { Buchanan & Duda 1983, p.191 } :

       Lsuf =  P( e| h)/P( e|~h) =   RR( e: h) = Qnec by I.J. Good
       Lnec =  P(~e| h)/P(~e|~h) = 1/RR(~e:~h) = [1 - P(e|h)]/[1 - P(e|~h)]
                                               = Qsuf by I.J. Good
   iff Lnec = 0     then  e  is logically necessary for h
   iff Lnec = large then ~e  is supporting h (ie absence of e supports h)

   Lsuf = Qnec , but in fact there is no semantic confusion, since
   Lsuf denotes how much is  e sufficient for h  (ie h necessary  for e), and
   Qnec denotes how much is  h necessary  for e  (ie e sufficient for h).


Folk2: { Brian Skyrms in James Fetzer, ed., 1988, p.172 }

       Ssuf = P( e| h)/P( e|~h) = RR( e| h) = Lsuf  interpreted as follows:
         iff  Ssuf > 1 then  h  has a tendency towards sufficiency for e

       Snec = P(~h|~e)/P(~h| e) = RR(~h:~e) = [1 - P(h|~e)]/[1 - P(h|e)]

       iff      Snec > 1 then  h  has tendency towards necessity for e
       iff Ssuf*Snec > 1 then  h  has tendency to cause the event e


Folk3:  { I.J. Good 1994, pp.306, + comment by P. Suppes on p.314 } :

       Qnec = P( e| h)/P( e|~h) = RR( e| h) = Lsuf , see Folk1 ;
       Qsuf = P(~e|~h)/P(~e| h) = RR(~e|~h) = [1 - P(e|~h)]/[1 - P(e|h)]
            = weight of evidence against h if e does not happen
            = a measure of causal tendency, according to I.J. Good.

       Iff Qnec*Qsuf > 1 then  h is a prima facie cause of e, adds Suppes.

       In { I.J. Good 1992, p.261 } his new insight is formulated thus :
  !!  "Qsuf(e:h) = Qnec(~e:~h). This identity is a generalization of the
       fact that h is a STRICT SUFFICIENT CAUSE of  e if and only if
                ~h is a STRICT NECESSARY  CAUSE of ~e, as any example
                 makes clear."     ( I.J. Good's emphasis )
  !!   Qnec(e:h) = Qsuf(~e:~h)  in { I.J. Good 1995, p.227 }
                 =   RR( e: h)  in { I.J. Good 1994, p.314 }, in my notation;
  !!   Qsuf(e:h) =   RR(~e:~h) is NOT ZERO if e,h are DISjoint, as my
                               common sense requires.

Folk4:  S = P(e|h) = " sufficiency of exposure h for effect e "
        N = P(h|e) = "   necessity of exposure h for effect e " { Schield }

  See Appendix A in { Schield 2002 } , his first lines left & right.
  There in his section 2.2 on necessity vs. sufficiency, Milo Schield
  nicely explains their contextual semantics and applicability thus:
  "Unless an effect [ e ] can be produced by a single sufficient cause [ h ]
   (RARE!), producing the effect requires supplying ALL of its necessary
   conditions [h_i], while preventing it [e] requires removing or eliminating
   only ONE of those necessary conditions." (I added the [.]'s ).
  Q: Well told, but do Schield's  S  and  N  fit his semantics ?
  A: No. While  S  is unproblematic, his  N  is not.
  Q: What does it mean that  h  is strictly sufficient for an effect e ?
  A: Whenever    h occurs, e  occurs too.  This in my formal translation
     means that  h implies e  ie  Peh = Ph  ie  P(e|h) = 1.
   Hence S = P(e|h)  measures sufficiency of h for e ,
                        or his  necessity of e for h (formally, I say).
   Note: if h =  bad exposure  and e = bad effect,    than all above fits;
         if h = good treatment for e = better health, than all above fits;
         other pairings would not fit meaningfully.

!! My view is this :  we are interested in  h CAUSES  e  (potentially).
       P(h|e) = Sufficiency of e for h  ie  e implies h ,
   or  P(x|y) = Sufficiency of y for x  ie  y implies x .
                Sufficiency is unproblematic, so I use it as a fundament.
  Q: What is  Nec = necessity of h for e , really ?
  A: I derive Nec from the semantical definition in { Schield 2002, p.1 }
     where he writes: "But epidemiology focuses more on identifying
  a necessary condition [h] whose removal would reduce undesirable outcomes
  [e] than on identifying sufficient conditions whose presence would produce
  undesirable outcomes."  His statement between [h] and [e] I formalize (by
  relying on the unproblematic Sufficiency ie on implication) thus:
  (no h) implies (no e)  ie  "no e without h" :
 ~h implies ~e, hence  P(~e|~h) =   1  in the  ideal extreme case.
   Note that generally P(~e|~h) = [ 1 - (Ph + Pe - Peh) ]/[ 1 - Ph ] = 1 here
!!      ie Peh = Pe ie  P(h|e)  =   1  in this IDEAL extreme case ONLY, while
                    N = P(h|e)  is Schield's general necessity of h for e.
!! But in general   N = P(h|e) <> P(~e|~h) which is [ see P(e|h) above ]
        SUFFICIENCY of  ~h for ~e, which better captures Schield's semantics.
 Q: Do we need his  N = P(h|e) ??
 A: Not if we stick to his more meaningful (than his N ) requirement on p.1
    just quoted, and opeRationalized by me thus :

!!! (Necessity of h for e) I define as (Sufficiency of ~h for ~e) == P(~e|~h)

  which is a COUNTERFACTUAL: IF no h THEN no e  ie "no e without h"
  which is close in spirit to I.J. Good's ( see Folk3 ) verbal definition,
                                except for the swapped suffixes nec and suf :
  Qsuf(e:h) = Qnec(~e:~h) = RR(~e:~h) in my notation as shown at Folk3 above.
 "Qsuf(e:h) = Qnec(~e:~h). This identity is a generalization of the
  fact that h is a STRICT SUFFICIENT CAUSE of  e  if and only if
           ~h is a STRICT NECESSARY  CAUSE of ~e, as any example
  makes clear." ( I.J. Good's emphasis; it took him 50 papers in 50 years ).

  The semantical  Necessity of h for e  as  removed h implies absence of e,
  is my  NecP = P(~e|~h), and not Schield's necessity N = P(h|e) not fitting
  his opeRational definition and containing no negation ~ as a COUNTERFACTUAL
  should.  Hence Schield's  N  is now deconstructed, and can be replaced by
  my constructive  NecP = P(~e|~h).
  Summary of  SufP = S , and of my NecP constructed from Schield's
!!!                          opeRationally meaningful verbal requirements :
  SufP = P( e| h)  = sufficiency of h for  e   defined as  h implies  e
  NecP = P(~e|~h)  =  necessity of  h for  e   defined as ~h implies ~e
  hence:
  SufP = P( e| h)  =  necessity of ~h for ~e           is  h implies  e

  Q: is my NecP ok ?
  A: Not yet, since for Peh = 0 my common sense requires S=0=N ie zero, which
     precludes all P(.,.) or P(.|.) containing ~ ie a NEGation.

     Fix1: IF Peh = 0 THEN  NecP1 = 0  ELSE  NecP1 = P(~e|~h).

     Fix2: Like Suf, Nec should have Peh as a factor in its numerator, so eg:
     SufP2 = P(e|h)   = SufP  = sufficiency of h for e  ie h implies e
     NecP2 = Peh*NecP = Peh*P(~e|~h) = Peh*P(~(e or h))/P(~h)
           = Peh*[ 1 - (Ph +Pe -Peh) ]/(1 - Ph)
           = necessity of h for e
     which seems reasonable, since without  Peh* my original  NecP  will be
     too often too close to 1 = max(P), hence poor anyway, as its form
     [ 1 - P(e or h) ]/[ 1 - Ph ]  is near 1 for small P's .


Folk5:  Jan Hajek's RR-formulas (derived as my criticism of Folk4):

   RRsuf = indication of how much the presence of h implies e
         = RR(h:e)  =     RR-sufficiency of h for e
         =  P(h|e)/P(h|~e)
         = [Peh/(Ph - Peh)] * (1 - Pe)/Pe                       (zzz)
         =  Peh*( h implies e) / Odds(e) , note that the "implies" factor
                  1/(Ph - Peh)  comes from P(h|~e), and that it is more
                          influential than P(h| e).

   RRnec = indication of how much absence of h implies absence of e
         = how much ~h implies ~e
         = derived from my "an event is an event is an event" & (zzz)
         =  RR(~h:~e)
         = [ P(~e,~h)/( P(~h) - P(~e,~h) )  ] * (1 - P(~e))/P(~e)
         = [ P(~e,~h)/          P( e,~h)    ] * Pe/P(~e)
         = [ P(~e,~h)/P(~e) ]/[ P( e,~h)/Pe ]
         =   P(~h|~e)/P(~h|e)
         = [1-P(h|~e)]/[1-P(h|e)]
!! RRnec should = 0 if Peh = 0, yet RRnec <> 0 here, but it will if we use
   RRnec*Peh  in analogy to NecP2 at the end of Folk4 .

 Since (h causes e) corresponds to (e implies h) we may have to swap e with h
 in some formulas above to get the (h causes  e). I have not always done it
 so to keep other authors' formulas as close to their original as reasonable.

.-

Finally lets look critically at the relation between causation and logical
implication.  "Rain causes us to wear a raincoat" does make sense, while
"NOT wearing a raincoat causes it NOT to rain" is an obvious NONSENSE, even
in a clean lab-like context with no shelter and our absolute unwillingness to
become wet.  Let x = rain  and  y = wearing a raincoat.
    The 1st statement translates to ( x causes  y );
    the 2nd statement translates to (~y causes ~x ).  Because nobody knows
how to formulate a perfect operator  "x causes  y", we substitute it with
"y implies x" (the swapped x, y is not the point, it doesnt matter just now).
```

```
Now the 1st statement translates to ( y implies  x );
    the 2nd statement translates to (~x implies ~y ).
But now we are in trouble, as  ( y implies x ) == ( ~x implies ~y ) in logic,
and ideally in probabilities :    ( Py = Pxy ) == ( P(~x) = P(~x,~y) )  ie:
in imperfect real situations :    ( Py - Pxy ) =  ( P(~x) - P(~x,~y) )  ie:
    Py - Pxy = (1 - Px) - (1 -(Px + Py - Pxy))
             =  1 - Px  -  1 + Px + Py - Pxy   =  Py - Pxy     q.e.d.
```

Hence such a simple difference doesnt work as we would like it did for a cause.

So what about the corresponding relative risks ? Lets check:

```
RR(y:x)          <>  RR(~x:~y)          ie they are not equal
 P(y|x)/P(y|~x)  <>   P(~x|~y)/P(~x|y)  ie:
(Pxy/(Py-Pxy))*((1-Px)/Px)  <>  ( (1-(Px+Py-Pxy))/(Py-Pxy) ) * (Py/(1-Py))
```

where we see the (y implies x) factor 1/(Py - Pxy) on both sides of the <> . Hence despite the <> both RR's will become oo ie infinite if (y implies x) perfectly whenever Pxy = Py. Otherwise, RR(y:x) is quite well behaved: RR increases with Pxy, and decreases with Px, which is reasonable as explained far above.

Conclusion: an implication cannot substitute causation in all its aspects, at least not in this case, hence not in general. But I dont know any other necessary (but not always sufficient) indicators of causal tendency than :

+ dependence (ie symmetrical association like eg correlation),
+ implication (ie asymmetrical association which is a transitive operation,
                  a subset in a subset in a subset in a subset ..etc).
      Caution: repeatedly find UNDESIRABLE above.
+ SurpriseBy
+ and time-ordering (a cause prior to the effect).
++ find key construction principles above for more and sharper formulations

-.-

+Acknowledgements (in alphabetic order) :

Jan Kahre of Aland & Helsinki, Finland, is an excellent discussion partner;
Leon Osinski of NL, is the best imaginable chief of a scientific library;
Mari Voipio of Helsinki is the best webmistress thinkable (multilingual too).

-.-

+References { refs } :

Q: Which refs are books and which are papers ?
A: Unlike in the titles of (e)papers here listed, in the titles of books and periodicals all words start with a CAP, except for the insignificants like eg.: a, and, der, die, das, for, from, in, of, or, to, the, with, etc.

Recalling Goethe's wisdom "In der Beschraenkung zeigt sich der Meister" I say: A too long list of refs would be almost as worthless as a too short list of refs. Therefore I did not include many (historically) relevant authors. If some are unreferred, then I mention one key group of German speaking Jewish emigre philosophers of scientific explanation/confirmation (& = co-op'ed) : C.G. Hempel & Paul Oppenheim & John G. Kemeny (Hungarian co-father of BASIC) & Olaf Helmer (he fathered the Delphi method of forecasting at RAND Corp.), Hans Reichenbach, Rudolf Carnap & Y. Bar-Hillel, Herbert Feigl, Kurt Grelling, and Karl R. Popper (Austrian) who played apart and refuted them all :-)

Computer Journal, UK, no.4, 1999, is a special issue on:
   - MML = minimum message length   (by Chris Wallace & Boulton, 1968)
   - MDL = minimum description length (by Jorma Rissanen, 1977)

- MLE = minimum length encoding (by Pendault, 1988)
These themes are very close to Kolmogorov's complexity, originated in the US by Occamite inductionists (as I call them) Ray Solomonoff in 1960, and Greg Chaitin in 1968.

Arendt Hannah: The Human Condition, 1959; on Archimedean point see pp.237 last line - 239, 260, more in her Index.

Agresti Alan: Analysis of Ordinal Categorical Data, 1984; see p.45 for a math-definition of Simpson's paradox for events A, B, C.

Agresti Alan: Categorical Data Analysis, 1st ed. 1990; see pp.24-25 & 75/3.24 on Goodman & Kruskal's TauB, Gini and Theil.

Agresti Alan: An Introduction to Categorical Data Analysis, 1996.

Alvarez Sergio A.: An exact analytical relation among recall, precision, and classification accuracy in information retrieval, 2002, see http://www.cs.bc.edu/~alvarez/APR/aprformula.pdf

Anderberg M.R.: Cluster Analysis for Applications, 1973.

Bailey N.T.J.: Probability methods of diagnosis based on small samples; in Mathematics and Computer Science in Biology and Medicine, Oxford 1964, 2nd printing 1966, pp.103-110

Blachman Nelson M.: Noise and Its Effects on Communication, 1966.

Blachman Nelson M.: The amount of information that y gives about X, IEEE Transactions on Information Theory, IT-14, Jan. 1968, 27-31

Blalock Hubert M.: Causal Inferences in Nonexperimental Research, 1964; start on p.62, on p.67 is his partial correlation coefficient.

Blalock Hubert M.: An Introduction to Social Research, 1970; on p.68 starts Inferring causal relationships from partial correlations.

Bar-Hillel Yehoshua: Language and Information, 1964, Addison-Wesley; the key paper is on pp.221-274. The Introductory chapter tells that its original & 1st author in 1952 was Rudolf Carnap.

Bar-Hillel Y., Carnap Rudolf: Semantic information, pp.503-511+512, in the book Communication Theory, 1953, Jackson W. (Willis) editor ; also in the British Journal for the Philosophy of Science, Aug. 1953. It is much shorter than the 1952 paper reprinted in Bar-Hillel, 1964, pp.221-274

Brin Sergey, Motwani R., Ullman Jeffrey. D., Tsur Shalom: Dynamic itemset counting and implication rules for market basket data, Proc. of the 1997 ACM SIGMOD Int. Conf. on Management of Data, 255-264. Sergey Brin is the co-founding father & CEO of Google, Inc.

Buchanan Bruce G., Duda Richard O.: Principles of rule-based expert systems; in Advances in Computers, vol.22, 1983, Yovits M.(ed).

Cheng Patricia W.: From covariation to causation: a causal power theory; (aka "power PC theory"), Psychological Review, 104 (1997), 367-405 = 39pp! Also see { Novick & Cheng 2004 }

Cheng Yizong, Kashyap Rangasami L.: A study of associative evidential reasoning, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.11, no.6, June 1989, pp. 623-631.

Cohen L. Jonathan: Knowledge and Language, 2002, Kluwer Academic Publishers;
! on p.180 in the eq.(13.8) both D should be ~D i.e. complements of D.

DeWeese M.R., Meister M.: How to measure the information gained from one symbol, Network: Computation Neural Systems 10, 1999, p.328.
They partially reinvented Nelson Blachman's fine work (see these refs).

Duda Richard, Gaschnig John, Hart Peter: Model design in the Prospector consultant system for mineral exploration; see pp.159 in { Michie 1979 }.

Eddy David M.: Probabilistic reasoning in clinical medicine: problems and opportunities, in Kahneman D., 1982, Judgment Under Uncertainty, 249-267.

Eells Ellery: Probabilistic Causality, 1991.

Fano Robert M.: Transmission of Information, 1961.

Feinstein Alvan R.: Principles of Medical Statistics, 2002, 701 pp; professor Feinstein, M.D. (1925-2001), at Yale (medicine), studied math & statistics; chap.10, pp.170-175 are on proportionate increments, on "excellent" NNE ie NNT, NNH; on honesty vs deceptively impressive magnified results.
Chap.17, pp.332,337-340 are on fractions, rates, ratios OR(:), risks RR(:).
! On p.340 the etiologic fraction should be e(r-1)/[e(r-1) +1] ;
! on p.444, eq.21.15 for negative likelihood ratio has swapped numerator and denominator: it should be (1-sensitivity)/specificity; above it should be (c/n1)/(d/n2) instead of the swapped ratio.

Fitelson Branden: Studies in Bayesian confirmation theory, Ph.D. thesis, University of Wisconsin-Madison, 2001, where I.J. Good's sinh should be tanh as I told him. Easily found on www.

Gigerenzer Gerd: Adaptive Thinking ; and his other fine books & papers.

Good I.J. (Irving Jack), born 1916 in London as "Isidore Jacob Gudak" who unlike Good is findable on WWW. He has been Alan Turing's main stats assistant during WWII when they were busily decoding other gentlemen's emails in Bletchley Park, UK. To understand my strange phrase "other gentlemen", you should know that a pre-WWII US Minister of War Henry Stimson has said that "Gentlemen do not read each other's mail". Tell it to you favorite 3-letter-word gov. agency :-) With Turing they were codebreaking secret codes produced by the German Enigma machine, of which some were passed to the British by the Polish resistance and Polish cryptologists who have done useful preanalysis (eg Rejewski).
I.J. has published over 1900 notes and papers (numbered 1-1900, and then not to confuse their numbers for dates, 2000-etc :-), of which some 50+ are on his favorite weight of evidence W(:).
! In my notation W(y:x) is his old W(x:y), and similarly with B(:), F(:).
Only since 1992 in his newest papers he switched to my notation (y:x).

Good I.J.: The mathematics of philosophy: a brief review of my work; in Critical Rationalism, Metaphysics and Science, 1995, Jarvie I.C. & Laor N. editors, pp.211-238.

Good I.J.: Legal responsibility and causation; pp.25-59 in the book Machine Intelligence 15, 1999, K. Furukawa, ed. Also see Michie in this volume 15.

Good I.J.: Causal tendency, necessitivity and sufficientivity: an updated review; pp.293-315 in "Patrick Suppes: Scientific Philosopher", vol.1, P. Humphreys, ed., 1994, Kluwer Academic Publ.
I.J.'s explains his fresh but surprisingly late insights (delayed 50 years) into the semantics of two W(:)'s, renamed by him to Qnec(y:x) , Qsuf(y:x), like mine ?(y:x)'s here, ie no more as his old W(x:y)'s.

On pp.312-315 Patrick Suppes' comments on I.J.'s paper.

Good I.J.: Tendencies to be sufficient or necessary causes, pp.261-262 in
Journal of Statistical Computation and Simulation, 44, 1992. This is a
preliminary note on Good's belated insight. 1992 - 1942 = 50 years
of delay. Delay is the deadliest form of denial :-).

Good I.J.: Speculations concerning the future of statistics, Journal
of Statistical Planning and Inference, vol. 25 (1990), 441-66.

Good I.J.: Abstract of "Speculations concerning the future of statistics",
The American Statistician, May 1990, vol. 44/2., 132-133.

Good I.J.: On the combination of pieces of evidence; Journal of Statistical
Computation and Simulation, 31, 1989, pp.54-58; followed by "Yet another
argument for the explicatum of weight of evidence" on pp.58-59.

Good I.J.: The interface between statistics and philosophy of science;
Statistical Science, 1988, vol.3, no.4, pp.386-412;
for W(:) see pp.389-390, 393-394 left low! + discussion & rejoinder p.409.

Good I.J.: Good Thinking - The Foundations of Probability and Its
Applications, 1983, University of Minnesota Pres. It reprints but a
fraction of his some 1500 papers and notes written until 1983.
On p.160 Kemeny & Oppenheim's degree of factual support F(:) is discussed;
! on p.160 up: I.J. Good's sinh(.) should be tanh(.).

Goodman Leo A., Kruskal William H.: Measures of Association for Cross
Classifications, 1979, 146 pp. Originally published under the same title
in the Journal of the American Statistical Association (JASA), parts 1-4:
part 1 in vol.49, 1954, pp.732-764; on TauB see pp.759-760
part 2 in vol.54, 1959, pp.123-163;
part 3 in vol.58, 1963, pp.310-364; on TauB see pp.353-354
part 4 in vol.67, 1972, pp. , on TauB see sect. 2.4
On ordinal measures see { Kruskal, 1958 } in JASA 53, 1958, pp.814-861.

Goodman Steven N.: Toward evidence-based medical statistics. Two parts:
1. The P value fallacy, pp. 995-1004, 2. The Bayes factor, pp.1005-1013;
discussion by Frank Davidoff: Standing statistics right up, pp.1019-1021;
all in Annals of Internal Medicine, 1999, very good, Goodman :-)

Grosof Benjamin N.: Evidential confirmation as transformed probability;
pp.153-166 in Uncertainty in Artificial Intelligence, L.N. Kanal
and J.F. Lemmer (editors), vol.1, 1986. I found that on p.159 his
! B == (1+C)/2 is in fact the rescaling in { Kemeny 1952, p.323 }, the
last two lines lead to F(:) rescaled on the first lines of p.324,
here & now findable as F0(:)

Grune Dick: How to compare the incomparable, Information Processing Letters,
24, 1987, 177-181.

Heckerman David R.: Probabilistic interpretations for MYCIN's certainty
factors; pp.167-196 in Uncertainty in Artificial Intelligence, L.N. Kanal
and J.F. Lemmer (eds), vol.1, 1986. I succeeded to rewrite his eq.(31) for
! the certainty factor CF2 on p.179 to Kemeny's F(:).
Heckerman has more papers in other volumes of this series of proceedings.

Hempel C.G.: Aspects of Scientific Explanation, 1965; pp.245-290 are chap.10,
Studies in the logic of explanation, reprinted from Philosophy of Science,
15 (reprinted paper of 1948 with Paul Oppenheim).

Hesse Mary: Bayesian methods; in Induction, Probability and Confirmation, 1975, Minnesota Studies in the Philosophy of Science, vol.6

Kac Mark: Enigmas of Chance, an autobiography, 1985.

Kahneman D. (ed): Judgment Under Uncertainty, 1982. He has won Nobel Prize
(economics, 2002) for this kind of work done with the late Amos Tversky.

Kahre Jan: The Mathematical Theory of Information, 2002, Kluwer Academic;
to find in his book formulas like eg Cont(.) use his special Index on
pp.491-493. See www.matheory.com or www.matheory.info for Errata + more.
! on p.120 eq(5.2.8) is P(x|y) - Px = Kahre's corroboration, x = cause,
! on p.186 eq(6.23.2) is P(y|x) - Py, risk is no corroboration; y = evidence

Kemeny John G., Oppenheim Paul: Degree of factual support; Philosophy of
Science, vol. 19, issue 4, Oct. 1952, 307-324. The footnote 1 on p.307
tells that Kemeny was de facto the author. Caution: on pp.320 & 324 his
! oldfashioned P(.,.) represents the modern P(.|.). On p.324 the first two
! lines should be bracketized and read P(E|H)/[P(E|H) + P(E|~H)], which is
findable here & now as F0(. An excellent paper worth of (y)our attention !

Kemeny John G.: A logical measure function, Journal of Symbolic Logic, 18/4,
December 1953, 289-308. On p.307 in his F(:) there are missing negation
! bars ~ over H's in both 2nd terms. Except for p.297 on Popperian elimination of models (find SIC here & now), there is no need to read this paper
if you read his much better one of 1952.

Kendall M.G., Stuart A.: The Advanced Theory of Statistics, 1977, vol.2.

Kruskal William H.: Ordinal measures of association, JASA 53, 1958, 814-861.

Lucas J.R., Hodgson P.E.: Spacetime and Electromagnetism, 1990;
see pp.5-13 on regraduation of speeds to rapidities.

Lusted L.B.: Introduction to Medical Decision Making, 1968.

Michie Donald: Adapting Good's Q theory to the causation of individual
events; pp.60-86 in Machine Intelligence 15, Furukawa K., Michie D. and
Muggleton S. (eds). During WWII at the age of 18, Michie was the youngest
codebreaker assisting I.J. Good who was Alan Turing's statistical assistant

Michie Donald (ed): Expert Systems in the Micro Electronic Age, 1979;

Norwich Kenneth: Information, sensation, and perception, 1993, Acad. Press.

Novick Laura R., Cheng Patricia W.: Assesing interactive causal influence;
Psychological Review, 111/2, 2004, pp.455-485 = 31 pages! Also see
{ Cheng Pat.W. 1997 }

Pang-Ning Tan, Kumar Vipin, Srivastava Jaideep: Selecting the right
interestingness measure for association patterns; kdd2002-interest.ps
is a comparative study of 21 measures of "interestingness"

Pearl Judea: Causality: Models, Reasoning, Inference, 2000; see at least
pp.284, 291-294, 300, 308; his references to Shep should be Sheps, and on
! p.304 in the Note under Tab.9.3 ERR = 1 - P(y|x')/P(y|x) would be correct

Popper Karl: Conjectures and Refutations, 1963, Routledge and Kegan Paul.

Popper Karl: The Logic of Scientific Discovery, 6th impression (revised),
March 1972; new appendices, Appendix IX (on corroboration) to his original
Logik der Forschung, 1935 (in his Index his Gehalt means SIC here & now).
His oldfashioned P(y,x) actually means modern P(y|x).

Renyi Alfred: New version of the probabilistic generalization of the large sieve, Acta Mathematica Academiae Scientiarum Hungaricae, vol. 10, 1959, 217-226, his correlation coefficient R between events is on p. 221 is also found in Kemeny & Oppenheim, 1952, p.314, eq.(7).

Renyi Alfred: Selected papers of Alfred Renyi, 1976, in 3 volumes.

Renyi Alfred: A Diary on Information Theory, 1987. The 3rd lecture discusses asymmetry and causality on pp.24-25+33.

Rescher N.: Scientific Explanation, 1970. See pp.76-95 for the chap.10 = The logic of evidence, where his Pr(p,q) actually means P(p|q). Very nice methodology of derivation, but the result is not spectacular :-)
! Note that on p.84 he suddenly switches from Pr(p|q) to Pr(q|p). Why ?

Romesburg H.C.: Cluster Analysis for Researchers, 1984.

Sackett David L., Straus Sharon, Richardson W. Scott, Rosenberg William, Haynes Brian: Evidence-Based Medicine - How to Practice EBM, 2nd ed, 2000. There is a Glossary of EBM terms, and Appendix 1 on Confidence intervals (CI), written by Douglas G. Altman of Oxford, UK.

```
!  Errors and typos reported by me:
   p. 73: in the footnote d/(c+d)    should be d/(b+d)
   p. 73: prevalence           32%   should be 31%
   p. 76:                    < 95    should be >= 95
   p. 79: in the nomogram 1000 & 0.001 should be moved toward their neighbours
   p. 80:                    CGPs    should be CPGs
   p.236: RRR = 1 - RR = 1 - p2/p1   should be RRR = 1 - RR = 1 - p1/p2
   p.237: Odds ratio's SE of logOR   should contain only + + + no - -
   p.238: specificity is b/(b+d)     should be d/(b+d)
   p.238: Table 3.5                  should be Table 3.3 (what is /82 ?? )
   p.239: likelihood ratios LR+ LR- SE are all wrong (several typos, results)
   p.250: journals journals          should be journals
   p.252: the first        increase  should be increases
   some 30+ typos are listed at  http://www.cebm.utoronto.ca/search.htm ,
   yet there is hope: the 3rd edition is in the works.
```

Schield Milo, Burnham Tom: Algebraic relationships between relative risk, phi and measures of necessity and sufficiency; ASA 2002; on www too. Find NAIVE , SIMPLISTIC here & now.

Schield Milo: Simpson's paradox and Cornfield's conditions; ASA 1999; on www too; an excellent multi-angle explanation of confounding, which is a very important subject, yet seldom & poorly explained in books on statistics. His section 8 can be complemented by reading { Agresti 1984, p45 } for a definition of Simpson's paradox for events A, B, C.

Schield Milo, Burnham Tom: Confounder-induced spuriousity and reversal for binary data: algebraic conditions using a non-iteractive linear model; 2003, on www too (slides nearby).

Shannon C.E., Weawer W.: The Mathematical Theory of Communication, 1949; different printings differ in page numberings. Here I refer to the 4th printing of the paperback edition, Sept. 1969, Univ. of Illinois Press.

Sheps Mindel C.: An examination of some methods of comparing several rates or proportions; Biometrics, 15 (1959), pp.87-97.

Shinghal R.: Formal Concepts in Artificial Intelligence, 1992; see chap.10 on Plausible reasoning in expert systems, pp.347-389, nice tables on
! pp.355-357, where in Fig.10.3 there are two typos in the necessity N which

should be $N = [1 - P(e|h)]/[1 - P(e|\sim h)]$;

! on p.352 just above 29. in the mid term (...) of the equation, both ~e should be e like in the section 10.2.11.

Simon Herbert: Models of Man, 1957. See pp.50-51 & 54.

Stoyanov J.M.: Counterexamples in Probability, 1987.

Suppes Patrick: A Probabilistic Theory of Causality, 1970.

Tversky Amos, Kahneman Daniel: Causal schemas in judgments under uncertainty; in Kahneman D.: Judgment Under Uncertainty, 1982, 117-128.

Vaihinger Hans: Die Philosophie des Als Ob - Volks-Ausgabe, 1923.

Van Rijsbergen C.J.: Information Retrieval, 2nd ed., 1979.

Weaver Warren: Science and Imagination, 19??; the section on "Probability, rarity, interest and surprise" has originally appeared in Scientific Monthly, LXVII ie 67, no.6, December 1948, pp.390 ??

Woodward P.M.: Probability and Information Theory, with Applications to Radar, 1953, Pergamon Press, 128 pages only. The 2nd edition of 1964 has 136 pages as it contains an additional chapter 8.

-.-