```
-- ## Stuff for tracking 'Num'bers.
-- "Num's track a list of number, and can report it sorted.
local Num-obj'*Num'
function Num.new(inits,at, txt, self)
self= has(Num, {at-at or 0, txt=txt or"", w=(txt or""):find"-" and -1 or 1,
has={1}, n=0, lo=1E32, hi =1E-32, ready=true})
for _,one in pairs(inits or {}) do self:add(one) end
return self end
- Ensure that the returned list of numbers is sorted.
unction Num:all(x)
if not self.ready then table.sort(self.has) end
self.ready = true
return self.has end
function Num:dist(a,b)
if a=="?" then b=self:norm(b); a = b>.5 and 0 or 1
elseif b=="?" then a=self:norm(a); b = a>.5 and 0 or 1
else a,b = self:norm(a), self:norm(b) end
return abs(a-b) end
-- Combine two 'num's.
function Num:merge(other, new)
new = Num.new(self.has)
for _, x in pairs(other.has) do new:add(x) end
return new end
-- Return a merged item if that combination
-- is simpler than its parts.
function Num:mergeable(other, new,b4)
new = self:merge(other)
b4 = (self.n*self.rsd() + other.n*other:sd()) / new.n
print("?"*,b4, new:sd(), new.n, self.n, other.n)
if b4 >= new:sd() then print("!"); return new end end
-- The 'mid' is the 50th percentile.

function Num:mid() return self:per(.5) end
-- Return 'x' normalized 0..1, lo.hi.
function Num:norm(x, lo,hi)
f x==""" then return x end
lo,hi = self.lo, self.hi
return abs(hi - lo) < 1E-32 and 0 or (x - lo)/(hi - lo) end
-- Return the 'p'-th percentile number.
function Num:per(p, t)
t = self:all()
      p = p^* \pm t / / 1

return \pm t < 2 and t [1] or t [p < 1] and 1 or p > \pm t and \pm t or p] end
 -- The 10th to 90th percentile range is 2.56 times the standard deviation. function Num:sd() return (self:per(.9) - self:per(.1))/ 2.56 end
fin.hi = math.huge
return spans end
-- ## Stuff for tracking 'Sym'bol Counts.
- 'Sym's track symbol counts and the 'mode' (most frequent symbol).
- local Sym=obj'Sym'
function Sym.new(inits,at,txt, self)
self has(Sym, |at-at or 0, txt=txt or "", has={}, n=0, mode=nil, most=0})
for _,one in pairs(inits or {}) do self:add(one) end
return self end
function Sym:add(x)
self.n = self.n + 1
self.has[x] = 1 + (self.has[x] or 0)
if self.has[x] > self.most then self.most, self.mode = self.has[x], x end end
 function Sym:dist(a,b) return a==b and 0 or 1 end
function Sym:mid() return self.mode end
-- Create one span holding row indexes associated with each symbol function Sym:spans(egs, xys.x)

xys = {}
for pos.eg in pairs(egs) do
    x = eg[self.at]
    if x -- "?" then
        xys[x] = xys[x] or {}
    push(xys[x], pos) end end
    return map(xys, function(x,t) return {lo=x, hi=x, has=Num(t)} end) end
-- ## Stuff for skipping all things sent to a column local Skip=obj"Skip" function Skip.newL_,at,txt) return has(Skip,{at=at or 0, txt=txt or"", n=0}) end function Skip:add(x) self.n = self.n + 1; return x end
```

```
-- Samples store examples. Samples know about
-- (a) lo,hi ranges on the numerics
-- and (b) what are independent 'x' or dependent 'y' columns.
local Sample = obj"Sample:
function Sample.new( src,self)
self = has(Sample, {names=nil, all={}, ys={}, xs={}, egs={}})
if src the
if type(src)=="sring" then for x in csv(src) do self:add(x) end end
if type(src)=="mible" then for _,x in pairs(src) do self:add(x) end end
return self end
  return serient
function Sample:add(eg, ako,what,where)
if not self.names
then—- create the column headers
self.names = eg
for at,x in pairs(eg) do
ako = x:find":" and Skip or x:match"^[A-Z]" and Num or Sym
what = push(self.all, ako({}), at, x))
if not x:find":" then
where = (x:find("+") or x:find("-")) and self.ys or self.xs
push(where, what) end end
else—- store another example; update column headers
push(self.egs, eg)
for at,x in pairs(eg) do if x ~= "?" then self.all[at]:add(x) end end
return self end
    function Sample:better(eq1,eq2,
    n,s1,s2,e = #self.ys, 0, 0, 2.71828

for__num_in pairs(self.ys) do
    a = num:norm(eq1[num.at])
    b = num:norm(eq2[num.at])
    s1 = s1 - e^(num.w * (a-b)/n)
    s2 = s2 - e^(num.w * (b-a)/n) end
    return s1/n < s2/n end</pre>
    function Sample:betters(egs)
  return sort(egs or self.egs, function(a,b) return self:better(a,b) end) end
   function Sample:clone( inits,out)
  out = Sample.new():add(self.names)
  for _,eg in pairs(inits or {})) do out:add(eg) ereturn out end
      function Sample:dist(eg1,eg2,
                                                                                                                                                          a,b,d,n,inc)
            -- Report mid of the columns function Sample:mid(cols) return lap(cols or self.ys,function(col) return col:mid() end) end
            - Return spans of the column that most reduces variance unction Sample:splitter(cols) function Worker(col) return self:splitter1(col) end return first(sort(lap(cols or sample.xs, worker), firsts))[2] end
   -- Return a column's spans, and the expected sd value of those spans.

function Sample:splitter1(col, spans, xpect)
spans = col:spans(self.egs)
--spans = lap(spans, shout)
--:xpect = sum(spans, function(_,span) return span.has.n*span.has:sd()/#self.egs end)
return (xpect, spans) end
                 Split on column with best span, recurse on each split.
nction Sample:tree(min, node,min,sub,splitter, splitter1)
  -- Split on column with best span, recurse on each split function Sampletree(min, node, min, sub, splitter, splitter) node = (node=self, kids={}) min = min or (#self.egs)^the.SMALL if #self.egs >= 2min then for _, span in pairs(self:splitter()) do sub = self:clone() for _, at in pairs(span.has) do sub:add(self.egs[at]) end push (node.kids, span) span.has = sub:tree(min) end end return node end
             - Find which leaf best matches an example 'eg'.
mction Sample:where(tree,eg, max,x,default)
if #kid.has==0 then return tree end
max = 0
            max = 0
for _,kid in pairs(tree.node) do
   if #kid.has > max then default,max = kid,#kid.has end
   x = eg[kid.at]
   if x ~= "" then
   if x <= kid.hi and x >= kid.lo then
        return self:where(kid.has.eg) end end end
return self:where(default, eg) end
            - Discretization tricks
- Input a list of {{x,y}..} values. Return spans that divide the 'x' values
- to minimize variance on the 'y' values.
unction div(xys, tiny, dull, merge)
function merge(04) -- merge adjacent spans if combo simpler to he parts
                    unction dav(xys, tiny, unit, merge)

local j, tmp = 0, ()
while j < 494 do
j = 649 do
j 
          span.hi = x
span.has:add(y) end
return merge(spans) en
```

```
| Section of the sect
```

```
the.eg={}
function the.eg.shuffle( t,u,v)
t={}
for i=1,32 do push(t,i) end
u = shuffle(copy(t))
v = shuffle(copy(t))
assert(#t == #u and u[1] ~= v[1]) end
function the.eg.lap()
assert(3==lap({1,2},function(x) return x+1 end)[2]) end
function the.eg.map()
assert(3==map({1,2},function(_,x) return x+1 end)[2]) end
function the.eg.tables()
  assert(20==sort(shuffle({{10,20},{30,40},{40,50}}),firsts)[1][2]) end
function the.eg.csv( n,z)
  n=0 for eg in the.csv(the.FILE) do n=n+1; z=eg end assert(n==399 and z[\#z]==50) end
function the.eg.rnds(     t)
    assert(10.2 == first(rnds({10.22,81.22,22.33},1))) end
function the.eg.binsym( s,col,tmp)

s=Sample(the.FILE)

col = s.all[4]
local function show(v) return out(rnds({v.n, v:mid(), v:sd()},0)) end

print(show(col))

tmp = s:splitter1(col)

for k,v in pairs(tmp[2]) do print(k,v.lo,v.hi,v.has.n, show(v.has)) end
end
```



```
local lib={}
  lib.Seed = 10019
- random integers
function lib.sandi(lo,hi) return math.floor(0.5 + lib.rand(lo,hi)) end
  -- random floats

function lib.rand(lo,hi, mult,mod)

lo, hi = lo or 0, hi or 1

lib.Seed = (16807 * lib.Seed) % 2147483647

return lo + (hi-lo) * lib.Seed / 2147483647 end
-- Table to string.

lib.cat = table.concat
-- Return a sorted table.

lib.sort = function(t, f) table.sort(t, f); return t end
-- Return first, second, last item.

lib.first = function(t) return [2] end
lib.second = function(t) return [2] end
-- Function for sorting pairs of items.

lib.firsts = function(s, b) return a[1] < b[1] end
-- Add to end, pull from end.

lib.pop = table.remove
lib.push = function(t, x) table.insert(t, x); return x end
  -- Random order of items in a list (sort in place).
function lib.shuffle(t, );
for i=t,2,-1 do j=lib.randi(1,1); t[i],t[j]=t[j],t[i] end; return t end
  -- Shallow copy function lib.copy(t, u) u={}; for k,v in pairs(t) do u[k]=v end; return u end
  \begin{array}{ll} \textbf{function} \ \ lib.top(t,n, & u) \\ u=\{\}; \textbf{for} \ \ k,v \ \ \textbf{in} \ \ pairs(t) \ \ \textbf{do} \ \ \textbf{if} \ \ k>n \ \ \textbf{then} \ \ \textbf{break} \ \ \textbf{end}; \ push(u,v) \ \ \textbf{end}; \ \ \textbf{return} \ \ u; \textbf{end} \\ \end{array} 
  --- Return a table's keys (sorted).
function lib.keys(t,u)
u={}
u={}
for k,_ in pairs(t) do if tostring(k):sub(1,1)~="_" then lib.push(u,k) end end
return lib.sort(u) end
     - Binary chop (assumes sorted lists)
unction lib.bchop(t,val,lt,lo,hi, mid)
lt = lt or function(x,y) return x < y end
lo,hi = lo or 1, hi or #t
while lo <= hi do
mid = (lo+hi) // 2
if lt(t[mid],val) then lo=mid+l else hi= mid-l end end
return math.min(lo,#t) end
 - Sum items, filtered through 'f'.
unction lib.sum(t,f)
f= f or function(x) return x end
out=0; for __x in pairs(f) do out = out + f(x) end; return out end
 EP RIJM TIN E
  lib.fmt = string.format
lib.say = function(...) print(lib.fmt(...)) end
  -- Print as red, green, yellow, blue.
function lib.color(s,n) return lib.color(s,31) end
function lib.green(s) return lib.color(s,31) end
function lib.yellow(s) return lib.color(s,32) end
function lib.blue(s) return lib.color(s,34) end
function lib.blue(s) return lib.color(s,36) end
  -- Printed string from a nested structure.
lib.shout = function(x) print(lib.out(x)) end
    t=()
for cell in line:gsub("[w]v]*",""):gsub("#*",""):gmatch("([^]+)") do
lib.push(t, tonumber(cell) or cell) end
line = io.read()
if #t>0 then return t end
else io.close(file) end end end
```

Dec 11, 21 13:30 Z.lua Page 2/2