



```

218 -----
219 --- CLASSES
220 ---
221 ---
222 Num, Sym, Egs = obj"Num", obj"Sym", obj"Egs"
223
224 ---
225 ---
226 ---
227
228 function Sym:new(at,name)
229     return new({at=at, name=name, most=0,n=0,all={}}, Sym) end
230
231 function Num:new(at,name)
232     return new({at=at, name=name, _all={}, w=(name or ""):find"$" and -1 or 1,
233                 n=0, sd=0, mu=0, m2=0, lo=math.huge, hi=-math.huge), Num) end
234
235 function Egs:new(names, i,col)
236     i = new({_all={}, cols={names=names, all={}, x={}, y={}}, Egs)
237     for at,name in pairs(names) do
238         col = push(i.cols.all, (name:find"^[A-Z]" and Num or Sym) (at,name) )
239         if not name:find"$" then
240             if name:find"$" then i.cols.class = col end
241             push(name:find"[+!]" and i.cols.y or i.cols.x, col) end end
242     return i end
243
244 ---
245 ---
246 ---
247
248 function Sym.copy(i) return Sym(i.at, i.name) end
249
250 function Num.copy(i) return Num(i.at, i.name) end
251
252 function Egs.copy(i,rows, j)
253     j = Egs(i.cols.names)
254     for _,row in pairs(rows or {}) do j:add(row) end
255     return j end
256
257 ---
258 ---
259 ---
260
261 function Egs.add(i,row)
262     push(i._all, row)
263     for at,col in pairs(i.cols.all) do col:add(row[col.at]) end end
264
265 function Sym.add(i,x,inc)
266     if x ~="?" then
267         inc = inc or 1
268         i.n = i.n+inc
269         i.all[x] = inc + (i.all[x] or 0)
270         if i.all[x] > i.most then i.most, i.mode = i.all[x], x end end end
271
272 function Sym.sub(i,x,inc)
273     if x ~="?" then
274         inc = inc or 1
275         i.n = i.n - inc
276         i.all[x] = i.all[x] - inc end end
277
278 function Num.add(i,x,_, d,a)
279     if x ~="?" then
280         i.n = i.n + 1
281         d = x - i.mu
282         i.mu = i.mu + d/i.n
283         i.m2 = i.m2 + d*(x - i.mu)
284         i.sd = (i.m2<0 or i.n<2) and 0 or ((i.m2/(i.n - 1))^0.5)
285         i.lo = math.min(x, i.lo)
286         i.hi = math.max(x, i.hi)
287         a = i._all
288         if #a < the.keep then i.ok=false; push(a,x)
289         elseif r() < the.keep/i.n then i.ok=false; a[r(#a)]=x end end end
290
291 function Num.sub(i,x,_, d)
292     if x ~="?" then
293         i.n = i.n - 1
294         d = x - i.mu
295         i.mu = i.mu - d/i.n
296         i.m2 = i.m2 - d*(x - i.mu)
297         i.sd = (i.m2<0 or i.n<2) and 0 or ((i.m2/(i.n - 1))^0.5) end end
298
299 ---
300 ---
301 ---
302
303 function Egs.better(i,row1,row2)
304     local s1, s2, n, a, b = 0, 0, #i.cols.y
305     for _,col in pairs(i.cols.y) do
306         a = col:norm( row1[col.at] )
307         b = col:norm( row2[col.at] )
308         s1 = s1 - 2.7183*(col.w * (a - b) / n)
309         s2 = s2 - 2.7183*(col.w * (b - a) / n) end
310     return s1 / n < s2 / n end
311
312 function Egs.bettors(i,j,k)
313     return i:better(j:mid(j.cols.all), k:mid(k.cols.all)) end
314
315 function Egs.mid(i,cols)
316     return map(cols or i.cols.y, function(col) return col:mid() end) end
317
318 function Num.mid(i) return i.mu end
319 function Sym.mid(i) return i.mode end
320
321 function Num.div(i) return i.sd end
322 function Sym.div(i, e)
323     e=0; for _,n in pairs(i.all) do
324         if n > 0 then e = e + n/i.n * math.log(n/i.n,2) end end
325     return -e end
326
327 function Num.norm(i,x)
328     return i.hi - i.lo < 1E-32 and 0 or (x - i.lo)/(i.hi - i.lo) end
329
330 function Num.all(i)
331     if not i.ok then table.sort(i._all); i.ok=true end
332     return i._all end

```

```

333 ---
334 ---
335 ---
336 function Num.dist(i,a,b)
337     if a=="?" and b=="?" then return 1 end
338     if a=="?" then b=i:norm(b); a=b<.5 and 1 or 0
339     elseif b=="?" then a=i:norm(a); b=a<.5 and 1 or 0
340     else a,b = i:norm(a), i:norm(b) end
341     return math.abs(a - b) end
342
343 function Sym.dist(i,a,b)
344     return a=="?" and b=="?" and 1 or a==b and 0 or 1 end
345
346 function Egs.dist(i,row1,row2, d)
347     d = sum(i.cols.x, function(c) return c:dist(row1[c.at], row2[c.at])^the.p end)
348     return (d/#i.cols.x)^(1/the.p) end
349
350 function Egs.dists(i,r1,rows)
351     return sort(map(rows,function(r2) return {i:dist(r1,r2),r2} end),firsts) end
352
353 function Egs.half(i, rows)
354     local project,far,some,left,right,c,lefts,rights
355     far = function(r,t) return per(i:dists(r,t), the.far)[2] end
356     project = function(r1, a,b)
357         a,b = i:dist(left,r1), i:dist(right,r1)
358         return {(a^2 + c^2 - b^2)/(2*c), r1} end
359     some = many(rows, the.some)
360     left = far(any(some), some)
361     right = far(left, some)
362     c = i:dist(left,right)
363     lefts,rights = i:copy(), i:copy()
364     for n,projection in pairs(sort(map(rows,project),firsts)) do
365         if n==#rows//2 then mid=row end
366         (n <= #rows//2 and lefts or rights):add( projection[2] ) end
367     return lefts, rights, left, right, mid, c end
368
369

```

```

360 -----
370 --- DISCRETIZE
371 ---
372 ---
373 ---
374 Bin=obj"Bin"
375 function Bin:new(col,lo,hi,n,div)
376   return new({col=col, lo=lo, hi=hi, n=n, div=div},Bin) end
377
378 function Bin.selects(i,row, x)
379   x = row[i.col.at]
380   return x=="?" or i.lo==i.hi and x==i.lo or i.lo<=x and x<i.hi end
381
382 function Bin.show(i)
383   if i.lo==i.hi then return fmt("%s=%s", i.col.name, i.lo) end
384   if i.lo==math.huge then return fmt("%s<%s", i.col.name, i.lo) end
385   if i.hi== math.huge then return fmt("%s>=%s",i.col.name, i.hi) end
386   return fmt("%s<=%s<%s", i.lo, i.col.name, i.hi) end
387
388 function Bin.distance2heaven(i, divs, ns)
389   return ((1 - ns:norm(i.n))^2 + (0 - divs:norm(i.div))^2)^0.5 end
390
391 --- discretize syms
392 ---
393 ---
394
395 function Sym.bins(i,j)
396   local xys= {}
397   for x,n in pairs(i.all) do push(xys, {x=x,y="left", n=n}) end
398   for x,n in pairs(j.all) do push(xys, {x=x,y="right",n=n}) end
399   return Bin:syms(i, Sym, xys) end
400
401 function Bin:syms(col, yclass, xys)
402   local out,all={}, {}
403   for _,xy in pairs(xys) do
404     all[xy.x] = all[xy.x] or yclass()
405     all[xy.x]:add(xy.y, xy.n) end
406   for x,one in pairs(all) do
407     push(out,Bin(col, x, x, one.n, one:div())) end
408   return out end
409
410 --- discretize nums
411 ---
412 ---
413
414 function Num.bins(i,j)
415   local xys, all = {}, Num()
416   for _,n in pairs(i._all) do all:add(n); push(xys,{x=n,y="left"}) end
417   for _,n in pairs(j._all) do all:add(n); push(xys,{x=n,y="right"}) end
418   return Bin:nums(i, Sym, sort(xys,function(a,b) return a.x < b.x end),
419     {#xys^the.minItems,
420      all.sd*the.cohen} end
421
422 function Bin:nums(col, yclass, xys, minItems, cohen)
423   local out,b4= {}, -math.huge
424   local function binsl(lo,hi)
425     local lhs, rhs, cut, div, xpect, xy = yclass(), yclass()
426     for j=lo,hi do rhs:add(xys[j].y) end
427     div = rhs:div()
428     for j=lo,hi do
429       lhs:add(xys[j].y)
430       rhs:sub(xys[j].y)
431       if lhs.n > minItems and rhs.n > minItems then
432         if xys[j].x ~ xys[j+1].x then
433           if xys[j].x - xys[lo].x > cohen and xys[hi].x - xys[j].x > cohen then
434             xpect = (lhs.n*lhs:div() + rhs.n*rhs:div()) / (lhs.n+rhs.n)
435             if xpect < div then
436               cut, div = j, xpect end end end end end
437           if cut
438             then binsl(lo, cut)
439                binsl(cut+1, hi )
440             else b4 = push(out, Bin(col, b4, xys[hi].x, #xys, div)).hi end
441           end
442         binsl(1,#xys)
443         out[#out].hi = math.huge
444         return out end

```

```

444 --- ><plain
445 ---
446 ---
447 ---
448 local xplain,xplans,selects,spanShow
449 function Egs.xplain(i,rows)
450   local stop,here,left,right,lefts0,rights0,lefts1,rights1
451   rows = rows or i._all
452   here = {all=rows}
453   stop = (#i._all)^the.minItems
454   if #rows >= 2*stop then
455     lefts0, rights0, here.left, here.right, here.mid, here.c = half(i, rows)
456     if #lefts0._all < #rows then
457       cuts = {}
458       for j,col in pairs(lefts0.col.x) do col:spans(rights0.col.x[j],cuts) end
459       lefts1,rights1 = {},{}
460       for _,row in pairs(rows) do
461         push(selects(here.selector, row) and lefts1 or rights1, row) end
462       if #lefts1 > stop then here.lefts = xplain(i,lefts1) end
463       if #rights1 > stop then here.rights = xplain(i,rights1) end end end
464       return here end
465
466 function xbestSpan(spans)
467   local divs,ns,n,div,stats,dist2heaven = Num(), Num()
468   function dist2heaven(s) return (((1 - n(s))^2 + (0 - div(s))^2)^.5,s) end
469   function div(s) return divs:norm( s.all:div() ) end
470   function n(s) return ns:norm( s.all.n ) end
471   for _,s in pairs(spans) do
472     add(divs, s.all:div())
473     add(ns, s.all.n) end
474   return sort(map(spans, dist2heaven), firsts)[1][2] end
475
476 function selects(span,row, lo,hi,at,x)
477   lo, hi, at = span.lo, span.hi, span.col.at
478   x = row[at]
479   if x=="?" then return true end
480   if lo==hi then return x==lo else return lo <= x and x < hi end end
481
482 function xplans(i,format,t,pre,how, sel,front)
483   pre, how = pre or "", how or ""
484   if t then
485     prepre or ""
486     front = fmt("%s%s%s",pre,how, #t.all, t.c and rnd(t.c) or "")
487     if t.lefts and t.rights then print(fmt("%-35s",front)) else
488       print(fmt("%-35s",front, o(rnds(mids(i,t.all),format))))
489     end
490     sel = t.selector
491     xplans(i,format,t.lefts, "| .. pre, spanShow(sel)..: ")
492     xplans(i,format,t.rights, "| .. pre, spanShow(sel,true) ..: ") end end

```

```

493 ---
494 ---
495
496 function quintiles(ts,width,  nums,out,all,n,m)
497 width=width or 32
498 nums=Num(); for _,t in pairs(ts) do
499     for _,x in pairs(sort(t)) do add(nums,x) end end
500 all,out = nums.all, {}
501 for _,t in pairs(ts) do
502     local s, where = {}
503     where = function(n) return (width*nums:norm(n))/1 end
504     for j = 1, width do s[j]="" end
505     for j = where(per(t,.1)), where(per(t,.3)) do s[j]="-" end
506     for j = where(per(t,.7)), where(per(t,.9)) do s[j]="-" end
507     s[where(per(t,.5))]= "-"
508     push(out,{display=table.concat(s),
509         data = t,
510         pers = map({.1,.3,.5,.7,.9},
511             function(p) return rnd(per(t,p))end)}) end
512
513 return out end
514
515 function smallfx(xs,ys,      x,y,lt,gt,n)
516 lt,gt,n = 0,0,0
517 if #ys > #xs then xs,ys=ys,xs end
518 for _,x in pairs(xs) do
519     for j=1, math.min(64,#ys) do
520         y = any(ys)
521         if y<x then lt=lt+1 end
522         if y>x then gt=gt+1 end
523         n = n+1 end end
524 return math.abs(gt - lt) / n <= the.cliffs end
525
526 function bootstrap(y0,z0)
527 local x, y, z, b4, yhat, zhat, bigger
528 local function obs(a,b, c)
529     c = math.abs(a.mu - b.mu)
530     return (a.sd + b.sd) == 0 and c or c/((x.sd^2/x.n + y.sd^2/y.n)^.5) end
531 local function adds(t, num)
532     num = num or Num(); map(t, function(x) add(num,x) end); return num end
533 y,z = adds(y0), adds(z0)
534 x = adds(y0, adds(z0))
535 b4 = obs(y,z)
536 yhat = map(y._all, function(y1) return y1 - y.mu + x.mu end)
537 zhat = map(z._all, function(z1) return z1 - z.mu + x.mu end)
538 bigger = 0
539 for j=1,the.boot do
540     if obs( adds(many(yhat,#yhat)), adds(many(zhat,#zhat))) > b4
541     then bigger = bigger + 1/the.boot end end
542 return bigger >= the.conf end
543
544 --- xxx mid has to be per and
545 --- XXXX implement same
546 --- XXX need tests for stats
547 function scottKnot(nums,      all,cohen)
548 local mid = function(z) return z.some:mid()
549 end
550 local function summary(i,j,      out)
551     out = copy(nums[i])
552     for k = i+1, j do out = out:merge(nums[k]) end
553     return out
554 end
555 local function div(lo,hi,rank,b4,      cut,best,l,l1,r,r1,nl,nor)
556     best = 0
557     for j = lo,hi do
558         if j < hi then
559             l = summary(lo, j)
560             r = summary(j+1, hi)
561             now = (l.n*(mid(l) - mid(b4))^2 + r.n*(mid(r) - mid(b4))^2) / (l.n + r.n)
562             if now > best then
563                 if math.abs(mid(l) - mid(r)) >= cohen then
564                     cut, best, l1, r1 = j, now, copy(l), copy(r)
565                 end end end
566             if cut and not l1:same(r1,the) then
567                 rank = div(lo, cut, rank, l1) + 1
568                 rank = div(cut+1, hi, rank, r1)
569             else
570                 for i = lo,hi do nums[i].rank = rank end end
571                 return rank
572 end
573 table.sort(nums, function(x,y) return mid(x) < mid(y) end)
574 all = summary(1,#nums)
575 cohen = all.sd * the.cohen
576 div(1, #nums, 1, all)
577 return nums end

```

```

578 -----
579 ---
580 ---
581 ---
582 ---
583 function go.last()
584     ok( 30 == last({10,20,30}, "lasts") end
585
586 function go.per( t)
587     t={};for i=1,100 do push(t,i*1000) end
588     ok(70000 == per(t,.7), "per") end
589
590 function go.many( t)
591     t={};for i=1,100 do push(t,i) end; many(t,10) end
592
593 function go.sum( t)
594     t={};for i=1,100 do push(t,i) end; ok(5050==sum(t), "sum")end
595
596 function go.sample( m,n)
597     m,n = 10^5,Num(); for i=1,m do n:add(i) end
598     for j=.1,.9,.1 do do push(t,i*j,m*0.05) end end
599     print(j,per(n:all(),j),ish(per(n:all(),j),m*j,m*0.05)) end end
600
601 function go.sym( s)
602     s=Sym(); map({1,1,1,1,2,2,3}, function(x) s:add(x) end)
603     ok(ish(s:div(),1.378, 0.001), "cnt") end
604
605 function go.num( n)
606     n=Num(); map({10, 12, 23, 23, 16, 23, 21, 16}, function(x) n:add(x) end)
607     print(n:div())
608     ok(ish(n:div(),5.2373, .001), "div") end
609
610 function go.nums( num,t,b4)
611     b4,t,num={}, {},Num()
612     for j=1,1000 do push(t,100*r(i)*j) end
613     for j=1,#t do
614         num:add(t[j])
615         if j%100==0 then b4[j] = fmt("%.5f",num:div()) end end
616         for j=#t,1,-1 do
617             if j%100==0 then ok(b4[j] == fmt("%.5f",num:div()), "div"..j) end
618             num:sub(t[j]) end end
619
620 function go.syms( t,b4,s,sym)
621     b4,t,sym, s={}, {},Sym(), "I have gone to seek a great perhaps."
622     t={}; for j=1,20 do s:gsub('\'',function(x) t[#t+1]=x end) end
623     for j=1,#t do
624         sym:add(t[j])
625         if j%100==0 then b4[j] = fmt("%.5f",sym:div()) end end
626         for j=#t,1,-1 do
627             if j%100==0 then ok(b4[j] == fmt("%.5f",sym:div()), "div"..j) end
628             sym:sub(t[j]) end
629     end
630
631 function go.loader( num)
632     for row in things(the.file) do
633         if num then num:add(row[1]) else num=Num() end end
634         ok(ish(num.mu, 5.455,0.001), "loadmu")
635         ok(ish(num.sd, 1.701,0.001), "loadsd") end
636
637 function go.egsShow( t)
638     oo(Egs{"name","Age","Weigh-"}) end
639
640 function go.egsHead( )
641     ok(Egs({"name","age","Weight!")).cols.x, "Egs") end
642
643 function go.egs( eggs)
644     eggs = csv2egs(the.file)
645     ok(ish(egs.cols.x[1].mu, 5.455,0.001), "loadmu")
646     ok(ish(egs.cols.x[1].sd, 1.701,0.001), "loadsd") end
647
648 function go.dist( ds,egs,one,d1,d2,d3,r1,r2,r3)
649     eggs = csv2egs(the.file)
650     one = eggs._all[1]
651     ds={};for j=1,20 do
652         push(ds,egs:dist(any(egs._all), any(egs._all))) end
653     oo(rnds(sort(ds),"%5.3f"))
654     for j=1,10 do
655         r1,r2,r3 = any(egs._all), any(egs._all), any(egs._all)
656         d1=egs:dist(r1,r2)
657         d2=egs:dist(r2,r3)
658         d3=egs:dist(r1,r3)
659         ok(d1<= 1 and d2 <= 1 and d3 <= 1 and d1>=0 and d2>=0 and d3>=0 and
660             eggs:dist(r1,r2) == eggs:dist(r2,r1) and
661             eggs:dist(r1,r1) == 0
662             d3 <= d1+d2, "dist"..j) end end
663
664 function go.far( eggs,lefts,rights)
665     eggs = csv2egs(the.file)
666     lefts, rights = eggs:half(egs._all)
667     oo(rnds(egs:mid()))
668     print(egs:betters(lefts, rights))
669     print(egs:betters(rights, lefts))
670     oo(rnds(lefts:mid()))
671     oo(rnds(rights:mid())) end
672
673 function go.bin( eggs,lefts,rights,cuts)
674     eggs = csv2egs(the.file)
675     lefts, rights = eggs:half(egs._all)
676     for n,col in pairs(lefts.cols.x) do
677         cuts= col:bins(rights.cols.x[n])
678         map(cuts,function(cut) print(col.name, cut.lo, cut.hi,cut.n, cut.div) end);
679     end end
680
681 the = settings(help)
682 go.main(the.todo, the.seed)
683 os.exit(go.fails)

```