


```

217 -----
218 --- CLASSES
219
220
221
222 Num, Sym, Egs = obj"Num", obj"Sym", obj"Egs"
223
224 --- create
225
226
227 function Sym:new(at,name)
228     return new({at=at, name=name, most=0,n=0,all={}}, Sym) end
229
230
231 function Num:new(at,name)
232     return new({at=at, name=name, _all={}, w=(name or ""):find"$" and -1 or 1,
233                 n=0, sd=0, mu=0, m2=0, lo=math.huge, hi=-math.huge), Num) end
234
235
236 function Egs:new(names, i,col)
237     i = new({_all={}, cols={names=names, all={}, x={}, y={} }, Egs)
238     for at,name in pairs(names) do
239         col = push(i.cols.all, (name:find"[A-Z]" and Num or Sym) (at,name) )
240         if not name:find"$" then
241             if name:find"$" then i.cols.class = col end
242             push(name:find"[+]"$ and i.cols.y or i.cols.x, col) end end
243     return i end
244
245 --- copy
246
247 function Sym.copy(i) return Sym(i.at, i.name) end
248
249 function Num.copy(i) return Num(i.at, i.name) end
250
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 --- update
258
259
260 function Egs.add(i,row)
261     push(i._all, row)
262     for at,col in pairs(i.cols.all) do col:add(row[col.at]) end end
263
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
273 function Sym.sub(i,x,inc)
274     if x ~= "" then
275         inc = inc or 1
276         i.n = i.n - inc
277         i.all[x] = i.all[x] - inc end end
278
279
280 function Num.add(i,x,_, d,a)
281     if x ~= "" then
282         i.n = i.n + 1
283         d = x - i.mu
284         i.mu = i.mu + d/i.n
285         i.m2 = i.m2 + d*(x - i.mu)
286         i.sd = (i.m2<0 or i.n<2) and 0 or ((i.m2/(i.n - 1))^0.5)
287         i.lo = math.min(x, i.lo)
288         i.hi = math.max(x, i.hi)
289         a = i._all
290         if #a < the.keep then i.ok=false; push(a,x)
291         elseif r() < the.keep/i.n then i.ok=false; a[r(#a)]=x end end end
292
293
294 function Num.sub(i,x,_, d)
295     if x ~= "" then
296         i.n = i.n - 1
297         d = x - i.mu
298         i.mu = i.mu - d/i.n
299         i.m2 = i.m2 - d*(x - i.mu)
300         i.sd = (i.m2<0 or i.n<2) and 0 or ((i.m2/(i.n - 1))^0.5) end end
301
302 ---
303 --- C/H/T/Y
304
305
306 function Egs.better(i,row1,row2)
307     local s1, s2, n, a, b = 0, 0, #i.cols.y
308     for _,col in pairs(i.cols.y) do
309         a = col:norm( row1[col.at] )
310         b = col:norm( row2[col.at] )
311         s1 = s1 - 2.7183*(col.w * (a - b) / n)
312         s2 = s2 - 2.7183*(col.w * (b - a) / n) end
313     return s1 / n < s2 / n end
314
315
316 function Egs.betters(i,j,k)
317     return i:better(j:mid(j.cols.all), k:mid(k.cols.all)) end
318
319
320 function Egs.mid(i,cols)
321     return map(cols or i.cols.y, function(col) return col:mid() end) end
322
323
324 function Num.mid(i) return i.mu end
325 function Sym.mid(i) return i.mode end
326
327
328 function Num.div(i) return i.sd end
329 function Sym.div(i, e)
330     e=0; for _,n in pairs(i.all) do
331         if n > 0 then e = e + n/i.n * math.log(n/i.n,2) end end
332     return -e end
333
334
335 function Num.norm(i,x)
336     return i.hi - i.lo < 1E-32 and 0 or (x - i.lo)/(i.hi - i.lo) end
337
338
339 function Num.all(i)
340     if not i.ok then table.sort(i._all); i.ok=true end
341     return i._all end

```

```

332 ---
333 --- class
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
344 function Sym.dist(i,a,b)
345     return a=="?" and b=="?" and 1 or a==b and 0 or 1 end
346
347
348 function Egs.dist(i,row1,row2, d)
349     d = sum(i.cols.x, function(c) return c:dist(row1[c.at], row2[c.at])^the.p end)
350     return (d/#i.cols.x)^(1/the.p) end
351
352
353 function Egs.dists(i,r1,rows)
354     return sort(map(rows,function(r2) return {i:dist(r1,r2),r2} end),firsts) end
355
356
357 function Egs.half(i, rows)
358     local project,far,some,left,right,c,lefts,rights
359     far = function(r,t) return per(i:dists(r,t), the.far) [2] end
360     project = function(r1, a,b)
361         a,b = i:dist(left,r1), i:dist(right,r1)
362         return ((a^2 + c^2 - b^2)/(2*c), r1) end
363     some = many(rows, the.some)
364     left = far(any(some), some)
365     right = far(left, some)
366     c = i:dist(left,right)
367     lefts,rights = i:copy(), i:copy()
368     for n,projection in pairs(sort(map(rows,project),firsts)) do
369         if n==#rows//2 then mid=row end
370         (n <= #rows//2 and lefts or rights):add( projection[2] ) end
371     return lefts, rights, left, right, mid, c end
372
373 ---
374 --- discretize
375
376
377 local bins,xbestSpan
378 function Sym.bins(i,j, out)
379     local xys,all,one,last,x,y,n = {}, {}
380     for x,n in pairs(i.all) do push(xys, {x=x,y="lefts", n=n}) end
381     for x,n in pairs(j.all) do push(xys, {x=x,y="rights",n=n}) end
382     for _,tmp in pairs(sort(xys,function(a,b) return a.x < b.x end)) do
383         x,y,n = tmp.x, tmp.y, tmp.n
384         if x ~= last then
385             last = x
386             one = push(all, {lo=x, hi=x, all=Sym(i.at,i.name)}) end
387             one.all:add(y,n) end
388     for _,cut in pairs(all) do
389         push(out, {col=col, lo=cut.lo, hi=cut.hi,
390                     n=cut.all.n, div=cut.all:div()}) end end
391
392
393 function Num.bins(i,j, out)
394     local xys, all = {}, Num()
395     for _,n in pairs(i._all) do all:add(n); push(xys, {x=n,y="left"}) end
396     for _,n in pairs(j._all) do all:add(n); push(xys, {x=n,y="right"}) end
397     bins(i, xys, (#xys)^the.minItems, all.sd^the.cohen, Sym, out) end
398
399
400 function bins(col, xys, minItems, cohen, yclass, out)
401     local tmp, b4 = {}
402     local function bins1(xys)
403         local lhs, rhs, cut, div = yclass(), yclass()
404         local function xpect(i,j) return (i.n*i:div() + j.n*j:div()) / (i.n+j.n) end
405         for _,xy in pairs(xys) do rhs:add(xy.y) end
406         div = rhs:div()
407         for j,xy in pairs(xys) do
408             lhs:add(xy.y)
409             rhs:sub(xy.y)
410             if lhs.n >= minItems and rhs.n >= minItems then
411                 if xy.x ~= xys[j+1].x then
412                     if xy.x - xys[j].x >= cohen and xys[#xys].x - xy.x >= cohen then
413                         if xpect(lhs,rhs) < div then
414                             cut, div = j, xpect(lhs,rhs) end end end end end
415                 then local upto,after = {},{}
416                 for n,xy in pairs(xys) do push(n<=cut and upto or after, xy) end
417                 bins1(upto)
418                 bins1(after)
419             else push(tmp, {col=col, lo=xys[1].x, hi=xys[#xys].x, n=#xys, div=div}) end
420         end
421     bins1(sort(xys, function(a,b) return a.x < b.x end))
422     if #tmp1 then
423         tmp[1].lo = -math.huge
424         tmp[#tmp].hi = math.huge
425         for _,bin in pairs(tmp) do
426             if b4 then bin.lo = b4.hi end
427             b4 = push(out,bin) end end end
428
429 ---
430 --- >explain
431
432
433 local xplain,xplans,selects,spanShow
434 function Egs.xplain(i,rows)
435     local stop,here,left,right,lefts0,rights0,lefts1,rights1
436     rows = rows or i._all
437     here = {all=rows}
438     stop = (#i._all)^the.minItems
439     if #rows >= 2*stop then
440         lefts0, rights0, here.left, here.right, here.mid, here.c = half(i, rows)
441         if #lefts0._all < #rows then
442             cuts = {}
443             for j,col in pairs(lefts0.col.x) do col:spans(rights0.col.x[j],cuts) end
444             lefts1,rights1 = {},{}
445             for _,row in pairs(rows) do
446                 push(selects(here.selector, row) and lefts1 or rights1, row) end
447             if #lefts1 > stop then here.lefts = xplain(i,lefts1) end
448             if #rights1 > stop then here.rights = xplain(i,rights1) end end end
449     return here end
450
451
452 function xbestSpan(spans)
453     local divs,ns,n,div,stats,dist2heaven = Num(), Num()
454     function dist2heaven(s) return {(1 - n(s))^2 + (0 - div(s))^2^.5,s} end
455     function div(s) return divs:norm( s.all:div() ) end
456     function n(s) return ns:norm( s.all.n ) end
457     for _,s in pairs(spans) do
458         add(divs, s.all:div())
459         add(ns, s.all.n) end
460     return sort(map(spans, dist2heaven), firsts) [1] [2] end
461
462
463 function selects(span,row, lo,hi,at,x)
464     lo, hi, at = span.lo, span.hi, span.col.at
465     x = row[at]
466     if x=="?" then return true end
467     if lo==hi then return x==lo else return lo <= x and x < hi end end
468
469
470 function xplans(i,format,t,pre,how, sel,front)
471     pre, how = pre or "", how or ""
472     if t then
473         pre=pre or ""
474         front = fmt("%s%s%s",pre,how, #t.all, t.c and rnd(t.c) or "")
475         if t.lefts and t.rights then print(fmt("%-35s",front)) else
476             print(fmt("%-35s",front, o(rnds(mids(i,t.all),format))))
477         end

```

```

468 sel = t.selector
469 xplains(i,format,t.lefts, "|".. pre, spanShow(sel)..":")
470 xplains(i,format,t.rights, "|".. pre, spanShow(sel,true) ..":") end end

471 ---
472 ---
473 function quintiles(ts,width,  nums,out,all,n,m)
474 width=width or 32
475 nums=Num(); for _,t in pairs(ts) do
476     for _,x in pairs(sort(t)) do add(nums,x) end end
477 all,out = nums.all, {}
478 for _,t in pairs(ts) do
479     local s, where = {}
480     where = function(n) return (width*nums:norm(n))/1 end
481     for j = 1, width do s[j]=" " end
482     for j = where(per(t,.1)), where(per(t,.3)) do s[j]="-" end
483     for j = where(per(t,.7)), where(per(t,.9)) do s[j]="-" end
484     s[where(per(t,.5))] = "|"
485     push(out,{display=table.concat(s),
486             data = t,
487             pers = map({.1,.3,.5,.7,.9},
488                     function(p) return rnd(per(t,p)) end)}) end
489
490 return out end
491
492 function smallfx(xs,ys,      x,y,lt,gt,n)
493 lt,gt,n = 0,0,0
494 if #ys > #xs then xs,ys=ys,xs end
495 for _,x in pairs(xs) do
496     for j=1, math.min(64,#ys) do
497         y = any(ys)
498         if y<x then lt=lt+1 end
499         if y>x then gt=gt+1 end
500     n = n+1 end end
501 return math.abs(gt - lt) / n <= the.cliffs end
502
503 function bootstrap(y0,z0)
504 local x, y, z, b4, yhat, zhat, bigger
505 local function obs(a,b, c)
506     c = math.abs(a.mu - b.mu)
507     return (a.sd + b.sd) == 0 and c or c/((x.sd^2/x.n + y.sd^2/y.n)^.5) end
508 local function adds(t, num)
509     num = num or Num(); map(t, function(x) add(num,x) end); return num end
510 y,z = adds(y0), adds(z0)
511 x = adds(y0, adds(z0))
512 b4 = obs(y,z)
513 yhat = map(y..all, function(y1) return y1 - y.mu + x.mu end)
514 zhat = map(z..all, function(z1) return z1 - z.mu + x.mu end)
515 bigger = 0
516 for j=1,the.boot do
517     if obs( adds(many(yhat,#yhat)), adds(many(zhat,#zhat))) > b4
518     then bigger = bigger + 1/the.boot end end
519 return bigger >= the.conf end
520
521 --- xxx mid has to be per and
522 --- XXX implement same
523 --- XXX need tests for stats
524 function scottKnot(nums,      all,cohen)
525 local mid = function(z) return z.some:mid()
526 end
527 local function summary(i,j,      out)
528     out = copy( nums[i] )
529     for k = i+1, j do out = out:merge(nums[k]) end
530     return out
531 end
532 local function div(lo,hi,rank,b4,      cut,best,l1,rl,r1,now)
533     best = 0
534     for j = lo,hi do
535         if j < hi then
536             l = summary(lo, j)
537             r = summary(j+1, hi)
538             now = (l.n*(mid(l) - mid(b4))^2 + r.n*(mid(r) - mid(b4))^2)
539                   / (l.n + r.n)
540             if now > best then
541                 if math.abs(mid(l) - mid(r)) >= cohen then
542                     cut, best, l1, rl = j, now, copy(l), copy(r)
543                 end end end end
544             if cut and not l1:same(rl,the) then
545                 rank = div(lo,      cut, rank, l1) + 1
546                 rank = div(cut+1, hi, rank, rl)
547             else
548                 for i = lo,hi do nums[i].rank = rank end end
549             return rank
550         end
551         table.sort(nums, function(x,y) return mid(x) < mid(y) end)
552         all = summary(1,#nums)
553         cohen = all.sd * the.cohen
554         div(1, #nums, 1, all)
555         return nums end

```

```

556 -----
557 ---
558 ---
559 ---
560
561 function go.last()
562   ok( 30 == last{10,20,30}, "lasts") end
563
564 function go.per( t)
565   t={};for i=1,100 do push(t,i*1000) end
566   ok(70000 == per(t,.7), "per") end
567
568 function go.many( t)
569   t={};for i=1,100 do push(t,i) end; many(t,10) end
570
571 function go.sum( t)
572   t={};for i=1,100 do push(t,i) end; ok(5050==sum(t), "sum")end
573
574 function go.sample( m,n)
575   m,n = 10^5,Num(); for i=1,m do n:add(i) end
576   for j=.1,.9,.1 do push(t,i*1000) end
577   print(j,per(n:all(),j),ish(per(n:all(),j),m*j,m*0.05)) end end
578
579 function go.sym( s)
580   s=Sym(); map({1,1,1,2,2,3}, function(x) s:add(x) end)
581   ok(ish(s:div(),1.378, 0.001), "ent") end
582
583 function go.num( n)
584   n=Num(); map({10, 12, 23, 23, 16, 23, 21, 16}, function(x) n:add(x) end)
585   print(n:div())
586   ok(ish(n:div(),5.2373, .001), "div") end
587
588 function go.nums( num,t,b4)
589   b4,t,num={}, {},Num()
590   for j=1,1000 do push(t,100*r() *j) end
591   for j=1,#t do
592     num:add(t[j])
593     if j%100==0 then b4[j] = fmt("%.5f",num:div()) end end
594   for j=#t,1,-1 do
595     if j%100==0 then ok(b4[j] == fmt("%.5f",num:div()), "div"..j) end
596     num:sub(t[j]) end end
597
598 function go.syms( t,b4,s,sym)
599   b4,t,sym, s={}, {},Sym(), "I have gone to seek a great perhaps."
600   t={}; for j=1,20 do s:gsub('.',function(x) t[#t+1]=x end) end
601   for j=1,#t do
602     sym:add(t[j])
603     if j%100==0 then b4[j] = fmt("%.5f",sym:div()) end end
604   for j=#t,1,-1 do
605     if j%100==0 then ok(b4[j] == fmt("%.5f",sym:div()), "div"..j) end
606     sym:sub(t[j]) end
607   end
608
609 function go.loader( num)
610   for row in things(the.file) do
611     if num then num:add(row[1]) else num=Num() end end
612     ok(ish(num.mu, 5.455,0.001), "loadmu")
613     ok(ish(num.sd, 1.701,0.001), "loads") end
614
615 function go.egsShow( t)
616   oo(Egs{"name", "Age", "Weigh-"}) end
617
618 function go.egsHead( )
619   ok(Egs({"name", "age", "Weight!"}).cols.x, "Egs") end
620
621 function go.egs( egs)
622   egs = csv2egs(the.file)
623   ok(ish(egs.cols.x[1].mu, 5.455,0.001), "loadmu")
624   ok(ish(egs.cols.x[1].sd, 1.701,0.001), "loads") end
625
626 function go.dist( ds,egs,one,d1,d2,d3,r1,r2,r3)
627   egs = csv2egs(the.file)
628   one = egs._all[1]
629   ds={};for j=1,20 do
630     push(ds,egs:dist(any(egs._all), any(egs._all))) end
631   oo(rnds(sort(ds), "%5.3f"))
632   for j=1,10 do
633     r1,r2,r3 = any(egs._all), any(egs._all), any(egs._all)
634     d1=egs:dist(r1,r2)
635     d2=egs:dist(r2,r3)
636     d3=egs:dist(r1,r3)
637     ok(d1<= 1 and d2 <= 1 and d3 <= 1 and d1>=0 and d2>=0 and d3>=0 and
638       egs:dist(r1,r2) == egs:dist(r2,r1) and
639       egs:dist(r1,r1) == 0
640       d3 <= d1+d2, "dist"..j) end end
641
642 function go.far( egs,lefts,rights)
643   egs = csv2egs(the.file)
644   lefts, rights = egs:half(egs._all)
645   oo(rnds(egs:mid()))
646   print(egs:betters(lefts, rights))
647   print(egs:betters(rights, lefts))
648   oo(rnds(lefts:mid()))
649   oo(rnds(rights:mid())) end
650
651 function go.bin( egs,lefts,rights,cuts)
652   egs = csv2egs(the.file)
653   lefts, rights = egs:half(egs._all)
654   for n,col in pairs(lefts.cols.x) do
655     cuts={}
656     col:bins(rights.cols.x[n],cuts)
657     map(cuts,function(cut) print(col.name, cut.lo, cut.hi) end); end end
658
659 -----
660 the = settings(help)
661 go.main(the.todo, the.seed)
662 os.exit(go.fails)

```