Pareto Charts Plus

Pareto Charts, improved Pareto Charts, and intercomparison of percentages

Giles Crane, MPH, ASA, NJPHA

New Jersey R User Group,
Sponsored by Mango Solutions,
also Central Jersey R User Group
October 22, 2013

References

- *Vilfredo Pareto (1896-7) .Cours d'économie politique professé à l'université de Lausanne*, 3 volumes.

- Juran, J.M. (1951). "The Economics of Quality" in *Quality Control Handbook.* ed. J.M. Juran, New York, McGraw-Hill.

- Leland Wilkinson (2006). Revising the Pareto Chart. The American Statistician, November, 60:4:332-334.

The "7 jewels in the Crown" (George Box), applied simply, well, and consistently, provide realistic quality control and improvement.

- Pareto Charts
- Cause & Effect Diagrams
- Control charts
- Check lists
- Scatter diagrams
- Histograms
- Graphs

Example 1: Patient falls in U. of Texas Health Science Center, Houston

71 reports of patient falls at the hospital

15 Attributed causes with frequencies

9 9 9 7 6 5 4 4 3 3 3 3 2 1

Leland Wilkinson's example.

Example 2: Illinois Asylum circa 1860 *Probable Causes of Insanity (*play: Mrs.Packard*).*

Miasmatic fevers	25
Indigestion	14
Religious anxiety	11
Loss of property	12
Epilepsy	12
Disappointed love	10
Puerperal fever	8
Masterbation	8
Domestic unhappiness	4
Spiritual rappings	4
Intemperance	4
Intense study	4
Unkind relatives	4
Loss of children	3
Blows on the head	2
Loss of wife	2
Excessive blood loss	2
Sudden cessation of bleeding from the nose	1
Paralysis	1
Concusion from steamboat explosion	1
Jealousy	1
Shipwreck	1
Fear of mob	1
Loss of husband	1
Jaundice	1
Scurvy	1
Slander	1
Unknown	56

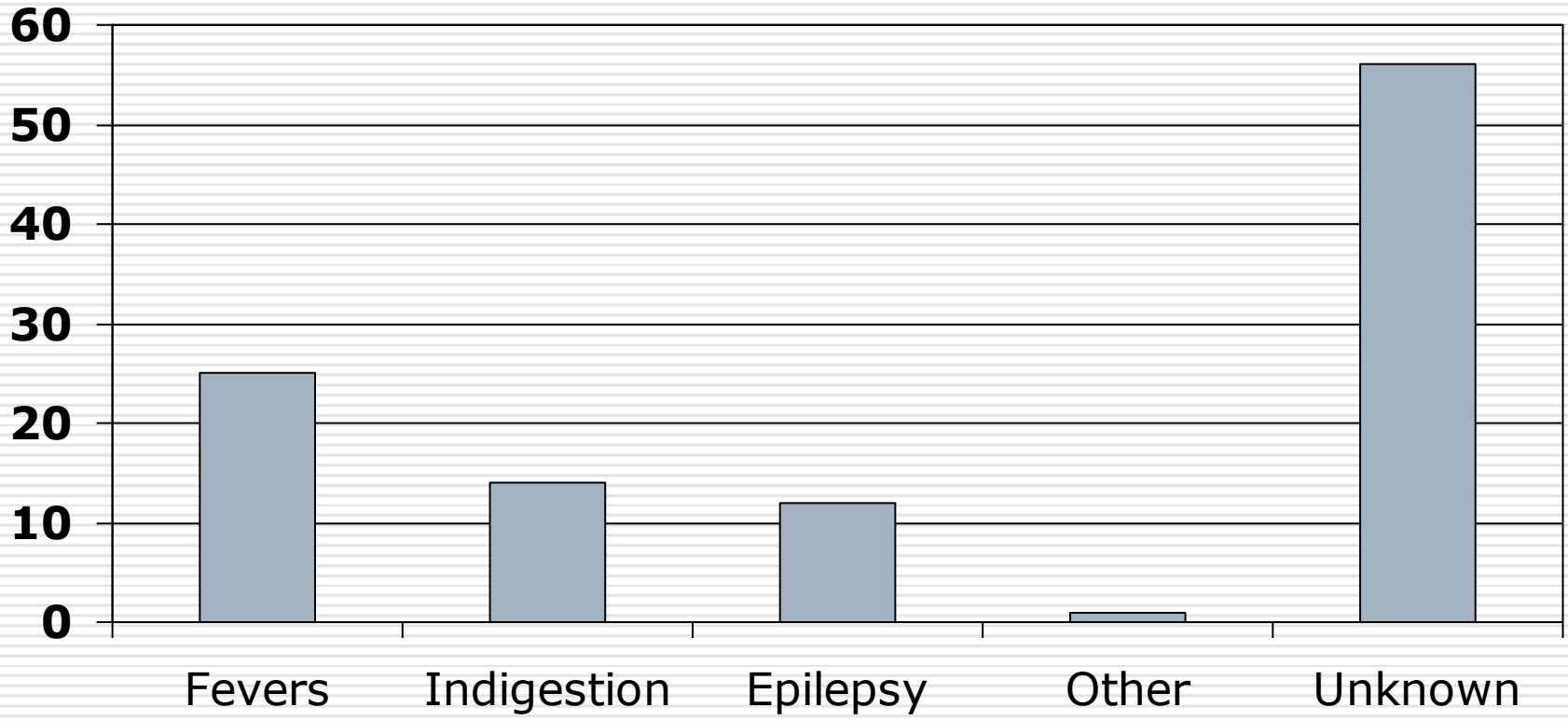

Jacksonville State Hospital
Illinois

EXAMPLE 3: Survey of HIV positive women: reasons for not seeking appropriate gynecological Care (NJ).

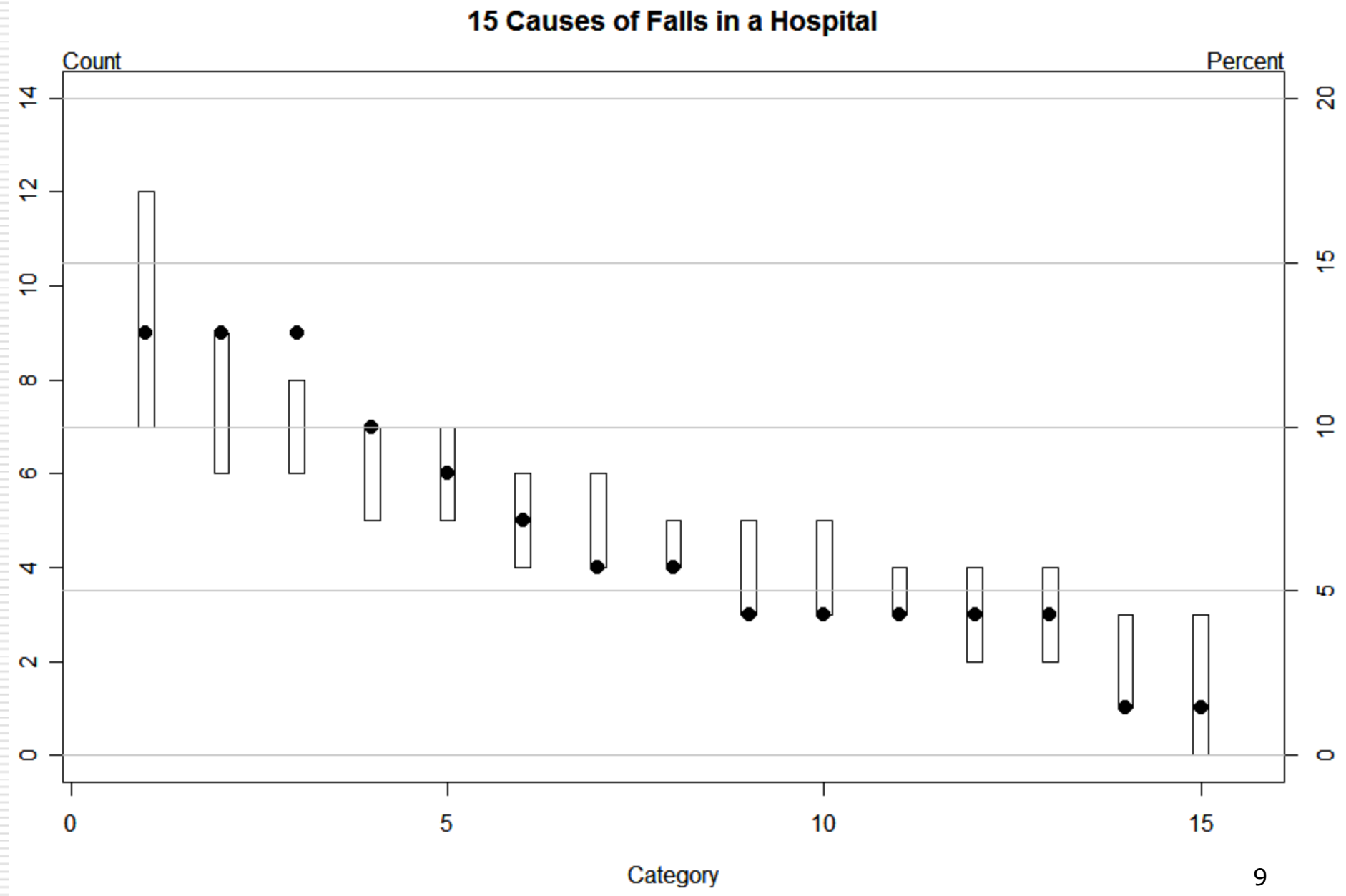
Fear of results	42
Procedural Discomfort	39
Cost	18
Competing Priorities	12
Appointment Time	10
Dislike Provider	6
Disclosure	2

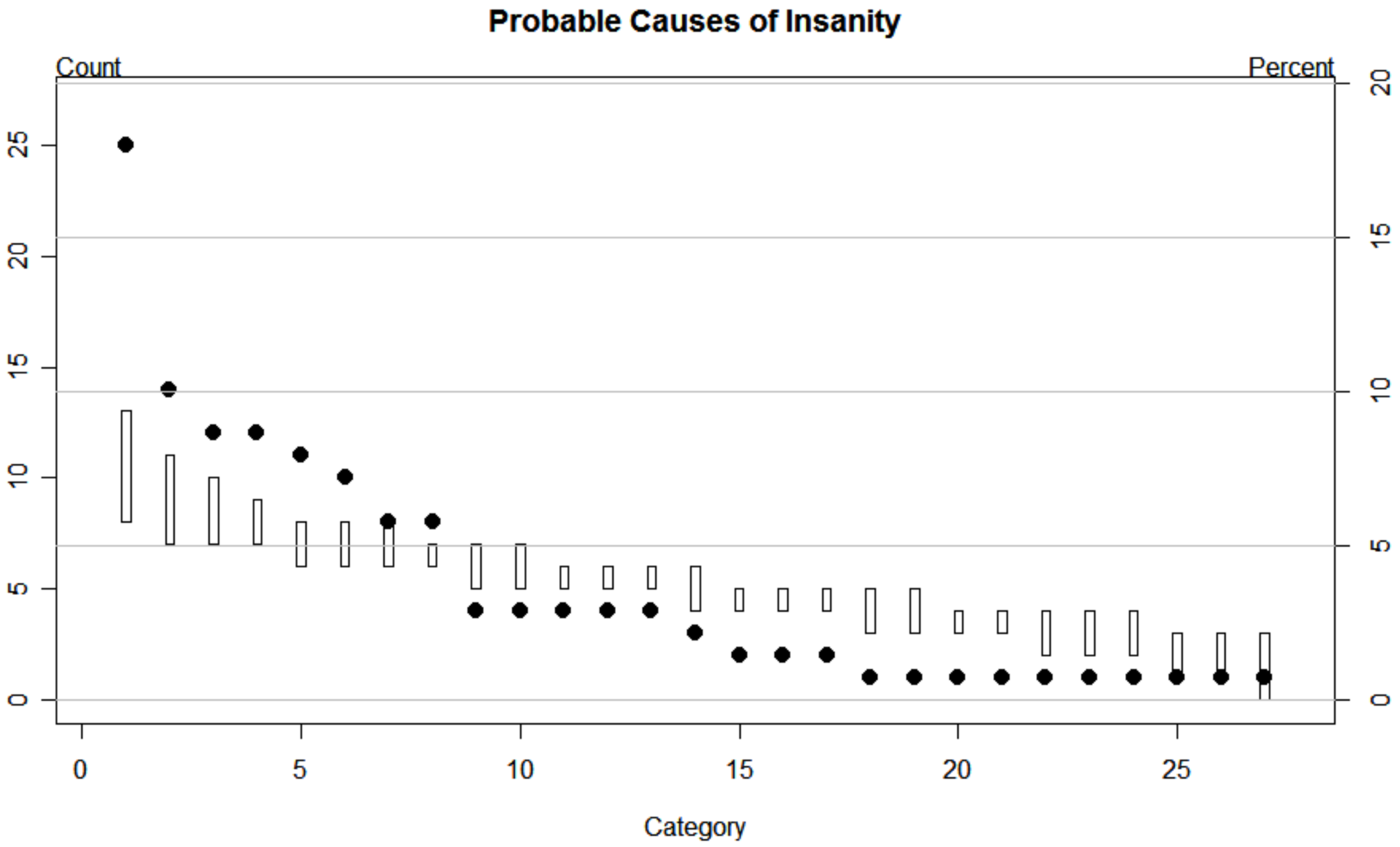
Courtesy of Jane Russo, NJ Dept. of Health

A Pareto Diagram is basically a bar chart of counts, arranged in decreasing order, but see the statistical limits in the next slides.

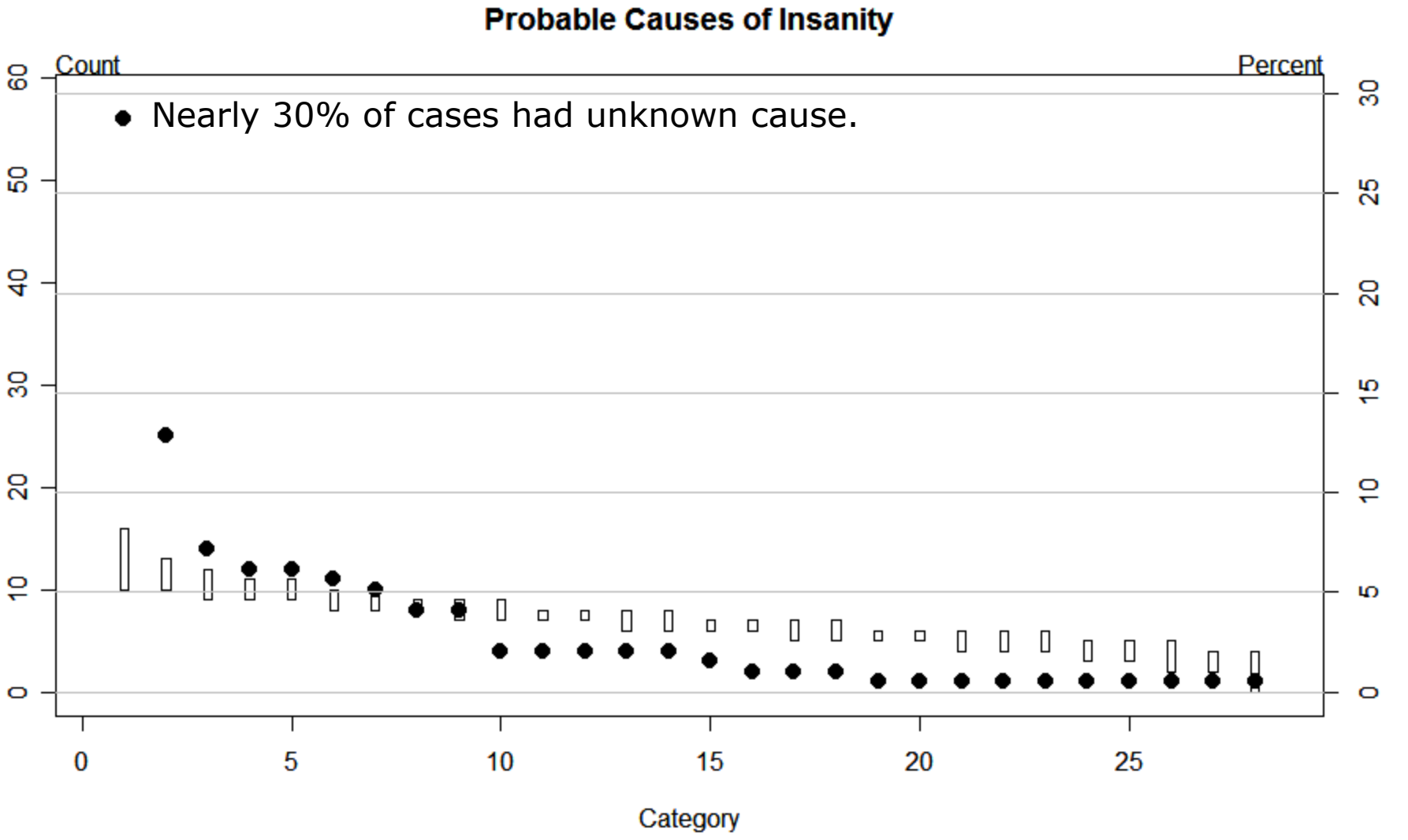


(Example 2) Causes of Insanity, abbreviated table.

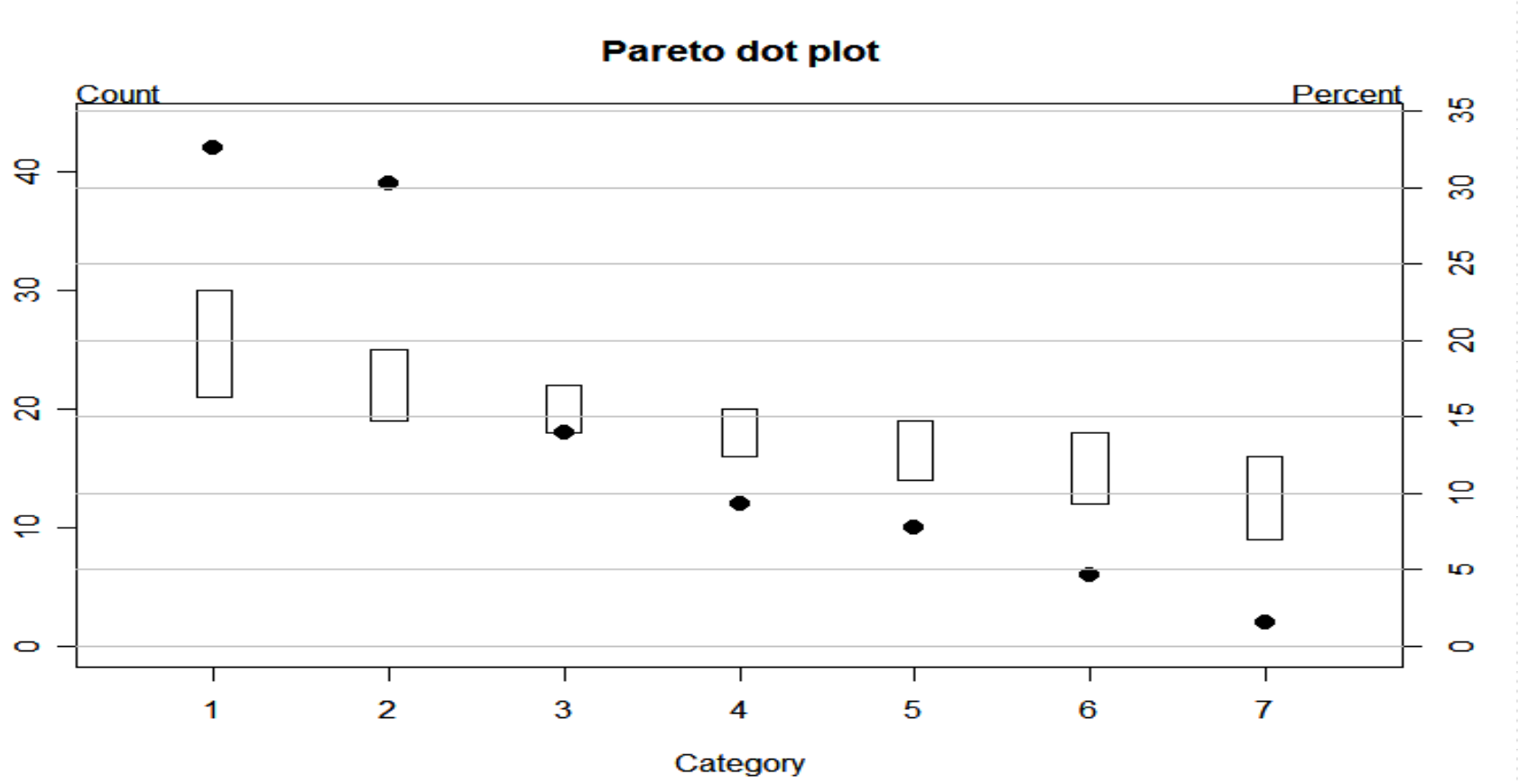

15 Causes of Falls in a Hospital
Count
Percent
0
2
4
6
8
10
12
14
0
5
10
15
20
0
5
10
15
Category


Probable Causes of Insanity
Count
Percent
0
5
10
15
20
25
Category

Including the unknown cause category, there are 28 categories.



With R, we can calculate limits indicating that counts within them may well be due to chance, not a real difference.



(Example 1) Reasons of HIV Positive women for not seeking gyn. care.

Statistical limits stimulate realistic considerations about the data.

- Do we have enough data?
- Difficulty of remedies for each cause?
- Cost for remedies?
- Side effects of remedies?
- Can certain causes be grouped?

In situations involving 3 or more counts, the Multinomial Distribution provides a theoretical background.

- P(n1, n2..nk | given p1, p2,..pk) = (N!/(n1! n2! .. nk!)) p1^n1 p2^n2..pk^nk
  where N = sum of ni

- Mean & variance of each ni are
  mean = N*pi
  var = N*pi*(1-pi)

But covariance ni and nj = -N pi pj

Acceptance Limits on multinomial counts in Pareto charts can be calculated by a computer simulation.

- Sample from multinomial distribution using number of categories, total count & equal probabilities, then sort counts, and save as a column.
- Do this several thousand times.
- For each row (thousands of columns), take the 2.5% count and the 97.5% count (quantiles), then plot as the lower and upper assurance limits.

Computing acceptance limits can be done by simulation in R. file pardot.r

```
# Obtain lower and upper acceptance limits on the ordered categories.

parlim  <- function(vn,  vp=c(0.025,0.975), nsample=2000) {

   ncat = length(vn)
   nsize = sum(vn)


   m = rmultinom(nsample, size = nsize, prob=rep(1,ncat) )


   dsort <-function(v) sort(v, decreasing=TRUE)
   m = apply(m, 2, dsort) # sort each column (Re-use m name)
```

Continue by taking quantiles and appending them to a data frame xlu. file pardot.r

```
# Obtain lower and upper assurance limits on the ordered categories.


    xlu = data.frame()        # Create buckets

    for (i in 1:ncat)  {
        limits = quantile(m[i,],vp)
        xlu = rbind(xlu, as.vector(limits))
        }

    row.names(xlu) = NULL     # Renumber rows 1,2,..#categories
    xlu = data.frame(vn,xlu)   # Put original counts as first column.
     return(xlu)
    }  #End of parlim function
```

We can now make a similar plot to that in Wilkinson's paper. File:pardot.r

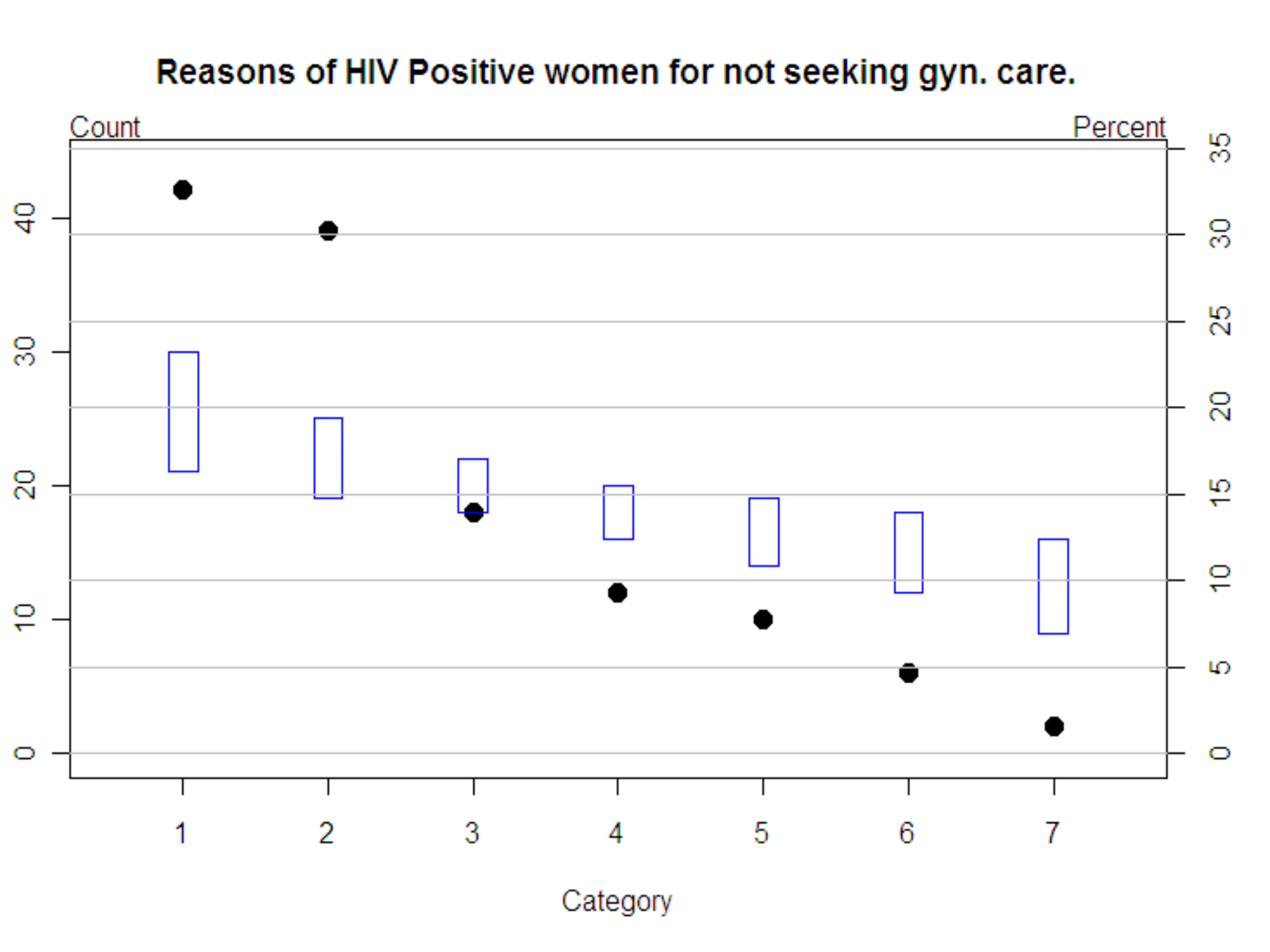
```
pardot <- function(xlu, catlabs,
                   heading="Pareto Dot Chart")  {
    ncat  = nrow(xlu)
    nsize = sum(xlu[,1])

    opar = par(mar=c(5,2,4,3) + 0.1 )
    plot(xlu[,1], main =  heading,
       xlab = "Category",
       ylab = "",
       xlim = c(0.5, ncat + 0.5),
       ylim = c(0, max(xlu[1,3],xlu[1,1]) + 2),

       pch = 19, cex=1.5)
```

Boxes are drawn around the points to dramatize the limits. File:pardot.r

```
# Assurance limits boxes
hw = 0.10  # Specify Half width for boxes
rect(1:ncat-hw, xlu[,2],1:ncat+hw,xlu[,3],
     border="blue")

ticks = axis(4,at=5*0:20*nsize/100,
             labels=as.character(5*0:20))
abline(h=5*0:20*nsize/100,col="gray")

mtext("Count",  side=3,line=0, adj=0)
mtext("Percent", side=3,line=0, adj=1)

par(opar)
}
```

Here is our NJ data example:

```
head1 = "Reasons of HIV Positive
   women for not seeking gyn. care."

vn=c(42,39,18,12,10,6,2)

xlu = parlim(vn)
pardot(xlu, ,head1)
```


Reasons of HIV Positive women for not seeking gyn. care.
Count
Percent
0
10
20
30
40
0
5
10
15
20
25
30
35
1
2
3
4
5
6
7
Category

The classical chi-square test also gives an overall test of equality of proportions: Are any pi different from any other?

- Chi-square Test: Consider Probability of outcome n1,n2,..nk , given p1=p2..=pk, i.e. equal probabilities.

For all k categories, Expected=n= N/k with equal probability.

Then

X = Sum of (ni – Expected)^2 / Expected.

is distributed as chi-square [sum of k N(0,1) variables]

If in tail of chi-square distribution, this is a rare event when all pi are equal.

Hence our sample may be from events of unequal pi.

Thus there are several, statistical approaches to answer: Are the percentages really different?

Omnibus Chi-square statistical test.

Alan Agresti's exact methods. (counts more extreme sets of count)

Simulation, including Pareto dot chart.

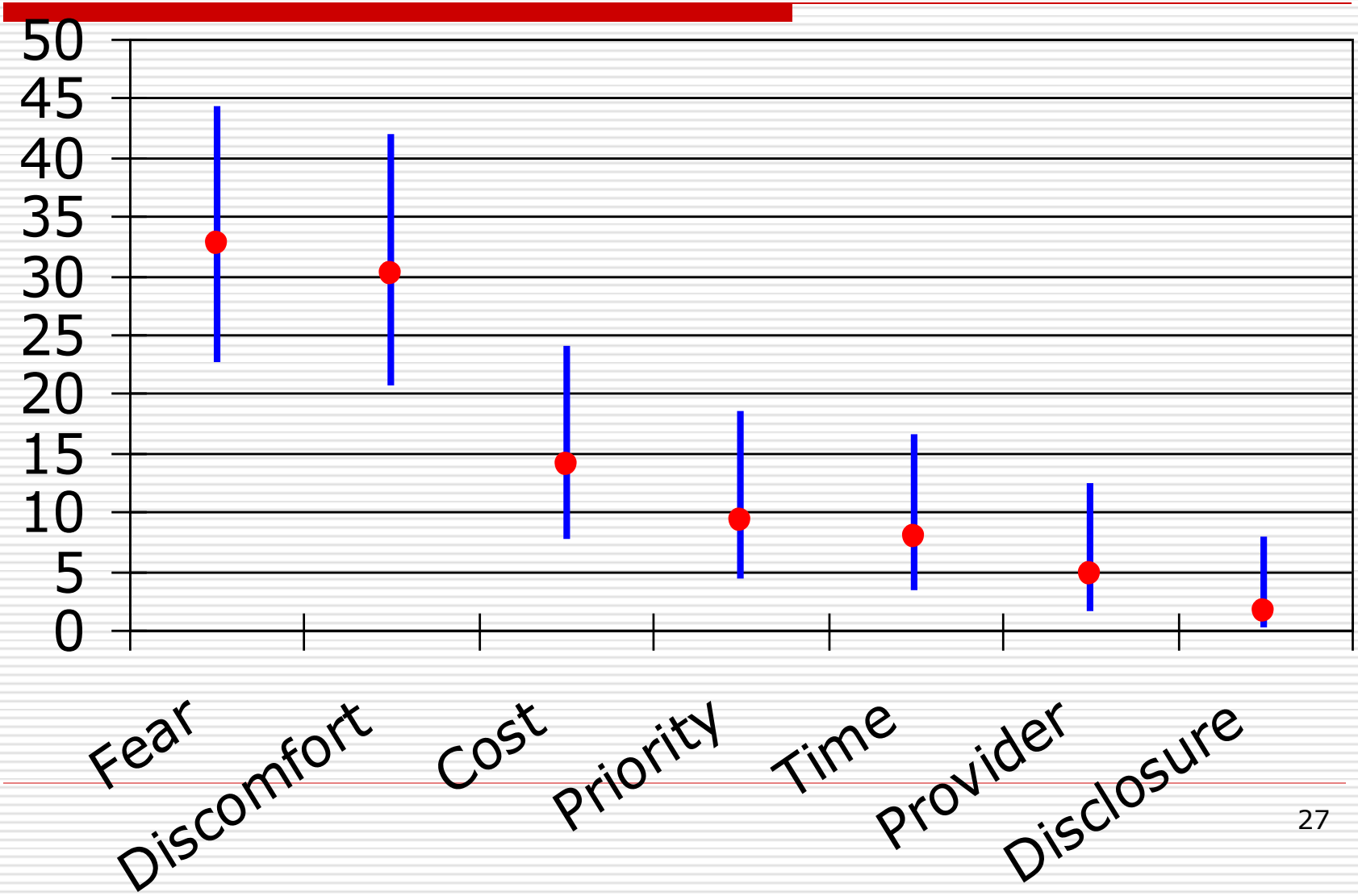
Inter-comparison of percentages.

Here are simulataneous confidence limits on the patient percentages for each reason HIV pos.women avoid proper care.

Estimate	Description	LCL95	UCL95
p1: 0.32558	Fear of results	0.226451	0.44324
p2: 0.30233	Procedural Discomfort	0.206457	0.41919
p3: 0.13953	Cost	0.076559	0.24081
p4: 0.09302	Competing Priorities	0.044293	0.18499
p5: 0.07752	Appointment Time	0.034370	0.16555
p6: 0.04651	Dislike Provider	0.016416	0.12478
p7: 0.01550	Disclosure	0.002859	0.07962

Fractions rather than percentages are shown in this table.

Statistical evidence for real differences can be seen in a plot of confidence intervals.



We computed simultaneous confidence limits in R. (multici.r)

Source of multici.r() Program: Mayo Foundation for Medical Education and Research (Ross Dierkhising, author)

Easily converted code from Splus to R

For one multinomial population or several multinomial populations.

Computes confidence limits on percentages, pairwise differences and contrasts.