Pearson, Kluyver and the Drunken Man

by: Michael A. Stephens

Department of Statistics and Actuarial Science
Simon Fraser University
Burnaby, British Columbia, Canada

stephens@stat.sfu.ca

Meeting to celebrate 1000000 Birthday of

Federico O'Reilly

UNAM, Mexico, D.F., Nov. 26, 2009

Un poco de historia

- November the 24th, 1859: Darwin's Origin of Species
- 1905: a Bumper year for Science
- 1945: WW2 ends, Mexico begins path to greater prosperity and world importance
- 1945, Music in Mexico: Ponce, coming to the end of a long career
- 1945, Art in Mexico: Diego Rivera, well known artist; Frida Kahlo and Rivera in 2nd marriage

Un poco de historia

- 1945: **Birth of a Mexican statistician**

1905: A bumper year for Science

- Physics: Einstein published three papers all of which had a tremendous impact

- 1905: Probability and statistics: Karl Pearson posed a random walk problem in Nature

1905: A bumper year for Science

- Physics: Einstein published three papers all of which had a tremendous impact

- 1905: Probability and statistics:
  Karl Pearson posed a random walk problem in Nature

- Pearson: A man starts at a point O and takes a step of one unit in any direction. He then takes a second step, at any randomly-oriented angle to the first, then a third at any angle, and so on

- R is the distance from O after n steps

- What is the distribution (or density) of R?

Correspondence in Nature with Lord Rayleigh

- Rayleigh: gave large-sample solution: ‘if n be very great, the probability sought is

$$(2/n)exp(-R^2/n)RdR'$$

- Nowadays: suppose X and Y are components of R on usual rectangular axes: $\sqrt{2/n}X$ and $\sqrt{2/n}Y$ asymptotically independent standard normal

- hence $2R^2/n = \chi^2_2$

Pearson has a sense of humour

- Pearson: 'The lesson of Lord Rayleigh's solution is that in open country the most probable place to find a drunken man who is at all capable of keeping on his feet is somewhere near his starting point!'

- This may be the first allusion to the drunken man.

Pearson has a sense of humour

- Pearson: 'The lesson of Lord Rayleigh's solution is that in open country the most probable place to find a drunken man who is at all capable of keeping on his feet is somewhere near his starting point!'

- This may be the first allusion to the drunken man.

- How many drunken men ever get to large n?

Help from Mr. Bennett

- But–Pearson says he wants the distribution for small n, but doesn't say why.

- Pearson also says he 'thanks Mr. Bennett for pointing out that for $n = 2$, solution is an elliptic integral'

- For $n = 2$, $f(R)$ is trivially easy: (and will be left to the reader to find)

Help from Mr. Bennett

- But–Pearson says he wants the distribution for small n, but doesn't say why.

- Pearson also says he 'thanks Mr. Bennett for pointing out that for $n = 2$, solution is an elliptic integral'

- For $n = 2$, $f(R)$ is trivially easy: (and will be left to the reader to find)

- **So: who was Mr. Bennett?**

- and can you prove the solution is an elliptic integral?

Kluyver to the rescue

- 1906: Kluyver gave a solution: distribution

$$F(R) = R \int_0^\infty \{J_0(t)\}^n \, J_1(Rt) \, dt$$

- Kluyver's elegant method allowed the steps to be of different lengths.

- Kluyver's method included in famous book on Bessel functions, Watson(1922).

- more of this later

Application: directional data

- directions (flights of birds, movements of insects re-acting to light, angles of pebbles)

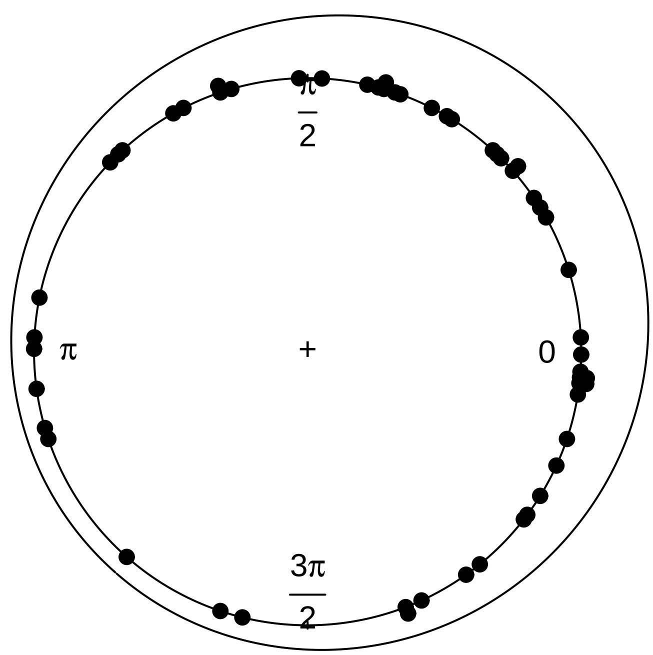
- expressed as vectors from centre O to points P around a unit circle; typical vector OP has angle θ

- The von Mises density :

$$f^*(\theta) = c\,\exp\{\kappa\cos(\theta - \theta_o)\};$$

- density $f_{vm}(R)$ of the vector sum of n von Mises vectors involves Kluyver's $f(R)$.

Plot of von Mises distribution for nematode data



Conditional tests of fit

- 1920's: Fisher's introduction of sufficiency vast amount of literature mostly on estimating parameters

- Lehmann (1950) famous book "Testing Hypotheses": gives conditions for optimal unbiased tests

- goodness-of-fit tests should be based on conditional distribution of data given sufficient statistics.

Federico's work on tests of fit

- Federico's early work; replace the usual PIT (using $z = F(x : \widehat{\theta})$) by $z = \tilde{F}(x|T)$, the Rao-Blackwell estimate of $F(x : .)$, given the sufficient statistic T

- recently (earlier with Rueda and later with Gracia-Medrano) directly create 'co-sufficient' samples

- co-sufficient samples: samples with the same sufficient statistic as the data

- Conditional tests for gamma distribution: 2008, with RAL and MAS

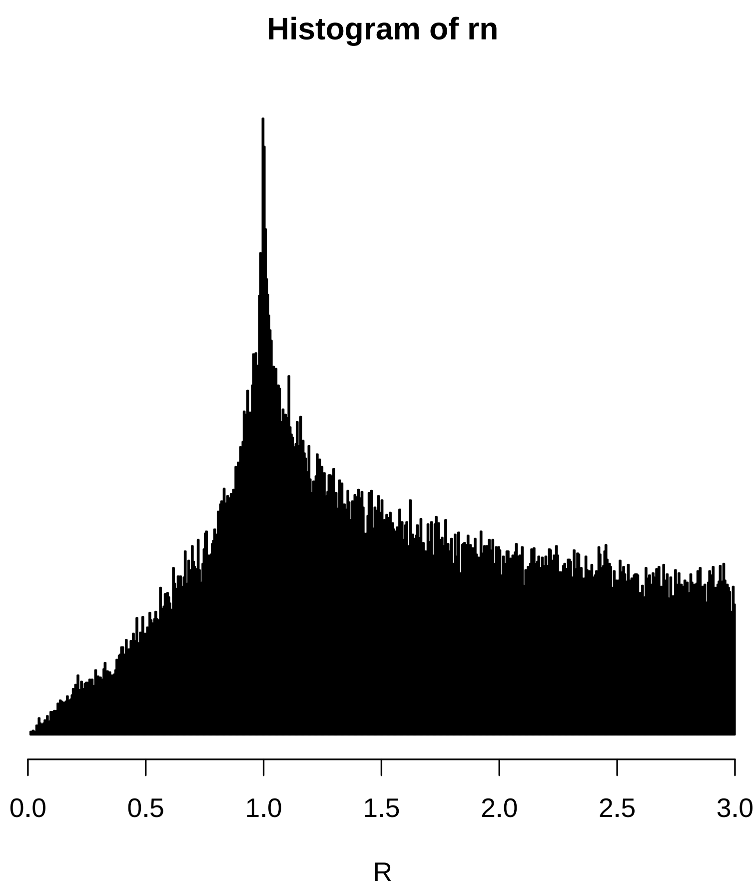
- innovation: Gibbs sampler used to get co-sufficient samples

Conditional tests for the von Mises distribution

- When using Gibbs sampler for tests for von Mises distribution we need Kluyver density $f(R)$ for $n = 3$.

- however, cannot differentiate $F(R)$ for $n < 4$

- Richard: huge Monte Carlo study to simulate $f(R)$ for $n = 3$, and finds huge spike at $R = 1$


Histogram of rn
0.0
0.5
1.0
1.5
2.0
2.5
3.0
R

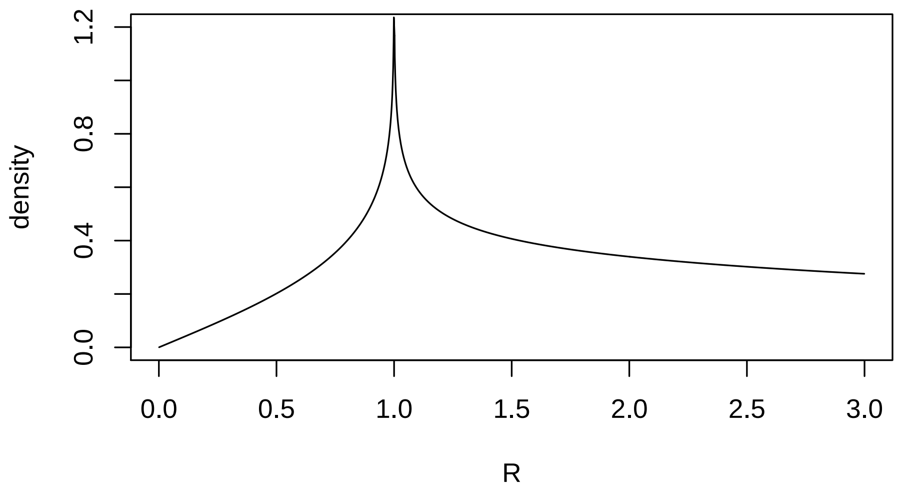
Michael gets a ride home

- Stephens (1962): gets density for $n = 3$ using elliptic integrals

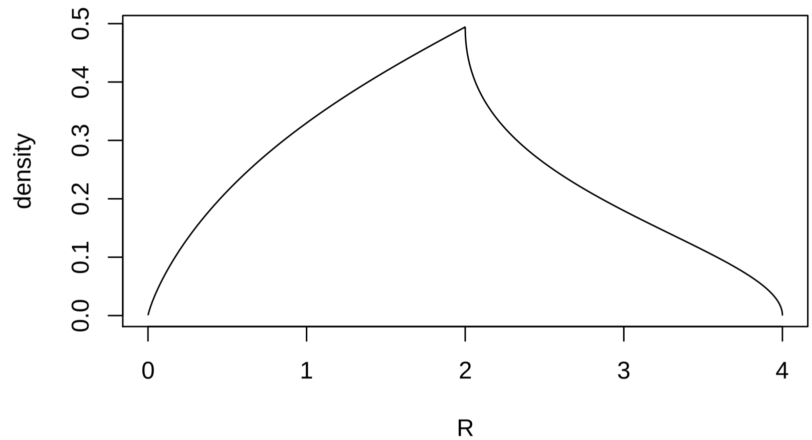
- when $R = 1$, density is infinite!
  interesting to wonder what types of walk give $R = 1$

- pictures show (strange?) densities, getting smoother with n

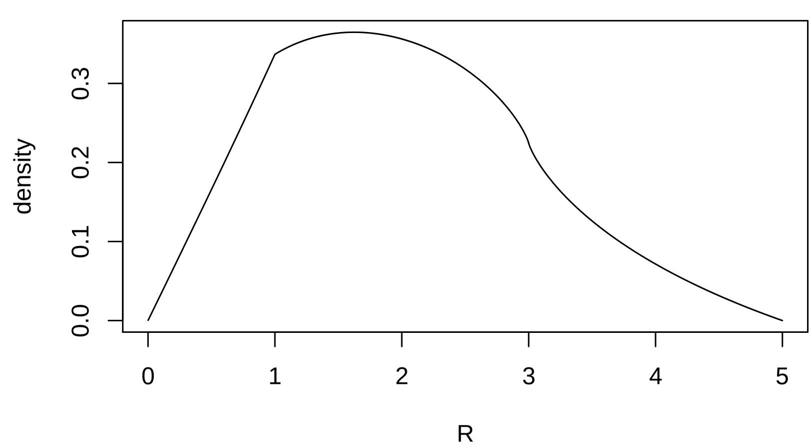
Density for $n = 3$



Density for $n = 4$



Density for $n = 5$



Wrapping it up

- Pearson's problem over 100 years ago gives fascinating results

- As $n \to \infty$, get Brownian motion

- can have drunken bees in 3D, drunken aliens in cyberspace

- recent article in Nature made drunken man walk on a lattice (Erdos posed problem) Have you ever seen one?

To be continued

- Here ends the fun part of the drunken man
- Richard continues tomorrow with the hard part

To be continued

- Here ends the fun part of the drunken man

- Richard continues tomorrow with the hard part (conditional tests)

Thanks to Federico

To be continued

- Here ends the fun part of the drunken man
- Richard continues tomorrow with the hard part (conditional tests)

Thanks to Federico

and

Thank you!

References

- Lehmann, E. L. and Romano, J. P. (2005) *Testing Statistical Hypotheses*, 3rd Edition. Springer, New York.

- Lockhart (2008a) Conditional limit laws for goodness-of-fit tests. *Manuscript*.

- Lockhart R. A., O'Reilly F. J. and Stephens M. A. (2007) Use of the Gibbs sampler to obtain conditional tests, with applications. *Biometrika*, **94**, 992–998.

O'Reilly, F. and Gracia-Medrano, L. (2004). On the conditional distribution of goodness-of-fit tests. *Serie Preimpresos*, **132**. Communicaciones Técnicos IIMAS, UNAM, México.

- O'Reilly, F. J. and Gracia-Medrano, L (2006) On the conditional distribution of goodness-of-fit tests. *Comms. Stats. (Theory and Methods)* **35**, 541–549.

  O'Reilly, F. and Rueda, R. (1992). Goodness of fit for the inverse Gaussian distribution. *The Canadian Journal of Statistics*, **20**, 387-397.

- Stephens M. A. (1962) Exact and approximate tests for directions. *Biometrika*, **49**, 463–477.