Statistics for HCI

# Part 1 – Wild and Wide

## randomness and distributions

Alan Dix

https://alandix.com/statistics/chi2022/



1

## unexpected wildness of random

just how random is the world?

raindrops and horse races



2

## *a story*

*In the far off land of Gheisra there lies the plain of Nali. For one hundred miles in each direction it spreads, featureless and flat, no vegetation, no habitation; except, at its very centre, a pavement of 25 tiles of stone, each perfectly level with the others and with the surrounding land.*

*The origins of this pavement are unknown – whether it was set there by some ancient race for its own purposes, or whether it was there from the beginning of the world.*

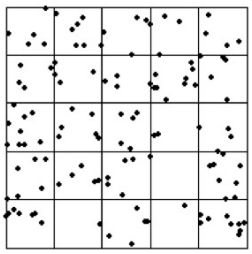
*Rain falls but rarely on that barren plain, but when clouds are seen gathering over the plain of Nali, the monks of Gheisra journey on pilgrimage to this shrine of the ancients, to watch for the patterns of the raindrops on the tiles. Oftentimes the rain falls by chance, but sometimes the raindrops form patterns, giving omens of events afar off.*

*Some of the patterns recorded by the monks are shown on the following pages. Which are mere chance and which foretell great omens?*



3

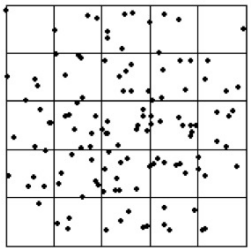
## *day 1*





4

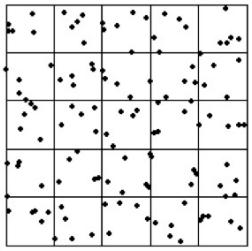
## *day 2*





5

## *day 3*

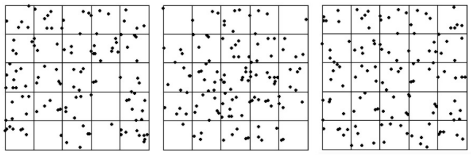




6

## your choice?

day 1 day 2 day 3



which are by chance ...
and which are unusual?



7

## why?

? .....

? .....

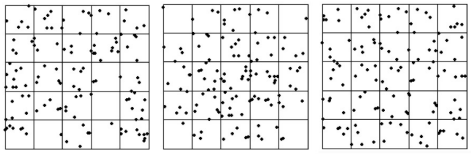
? .....



8

## which did you choose?

day 1 day 2 day 3

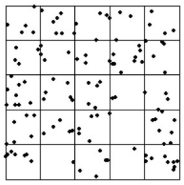


chance or omen?



9

## day 1 – really random

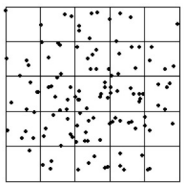


empty squares & overfull squares



10

## day 2 – random but not uniform

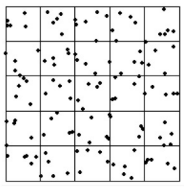


clumped towards the middle



11

## day 3 – too uniform



every square has 5 rain drops
too good to be true!



12



13

## two horse races

toss 20 coins

add the heads to one row
the tails to a second

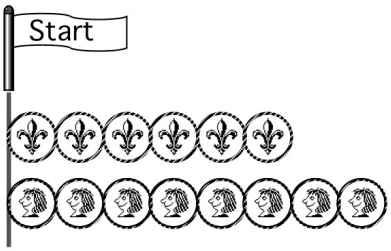
the winner is the first row to 10

before you start:
what do you think will happen?



14

## the race




did you get a clear winner?
or was it neck and neck?



15

## the world is very random

probability head = 0.5

number of heads $\neq \frac{\text{number of tosses}}{2}$



16

## lessons

apparent differences may be chance

real data has some bad values

e.g. Mendel's sweet peas and electron charge discovery were both too good!



17



18

## distributions

discrete or continuous
bounds and tails



19

## what kind of data

continuous:

e.g. time to complete task (12.73 secs)

discrete ...

arithmetic (interval):

e.g. number of errors (average makes some sense)

ordered/ordinal:

e.g. satisfaction rating (?average rating?)

nominal/categorical:

e.g. menu item chosen ( (File+Font)/2 = FIml ?)



20

## finite or unbounded

number of heads in 6 tosses

discrete, finite

number of heads until first tail

discrete, unbounded

wait before next bus

continuous, bounded below (zero), … but not above!

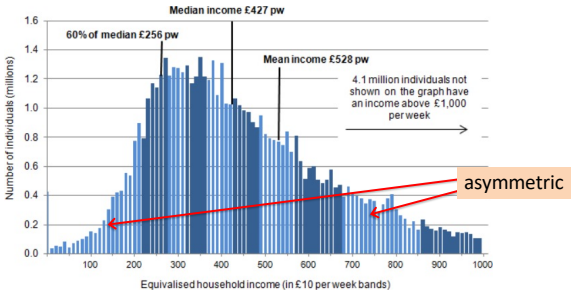
difference between heights

continuous, (sort of) unbounded



21

## distribution graph (e.g. UK income 2011/12)

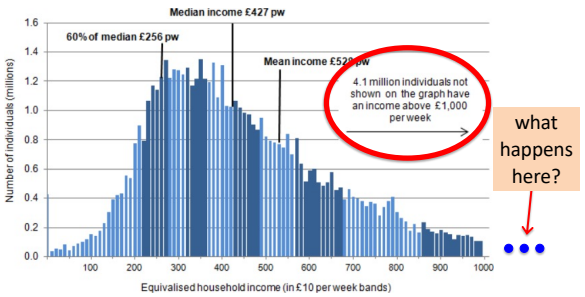



From: Sustainable Development Indicators, Office of National Statistics, July 2014
http://webarchive.nationalarchives.gov.uk/20160105183323/http://www.ons.gov.uk/ons/rel/wellbeing/sustainable-development-indicators/july-2014/sustainable-development-indicators.html



22

## long tail

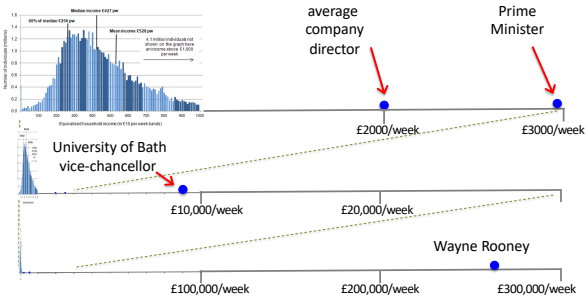



From: Sustainable Development Indicators, Office of National Statistics, July 2014
http://webarchive.nationalarchives.gov.uk/20160105183323/http://www.ons.gov.uk/ons/rel/wellbeing/sustainable-development-indicators/july-2014/sustainable-development-indicators.html



23

## long tail (ctd)




From: Sustainable Development Indicators, Office of National Statistics, July 2014
http://webarchive.nationalarchives.gov.uk/20160105183323/http://www.ons.gov.uk/ons/rel/wellbeing/sustainable-development-indicators/july-2014/sustainable-development-indicators.html



24

## one or two tails, …

what is your question?
do you care which direction?

is error rate *higher*?
- one tailed (discrete)

are completion times *different*?
- two tailed (continuous)



25



26

## normal or not

approximations
central limit theorem
power law



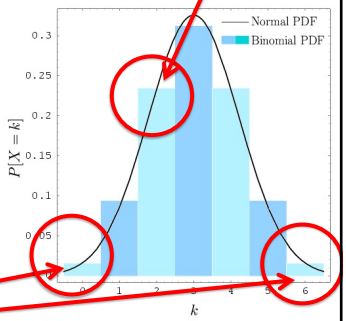
27

## approximations

continuous / discrete

may approximate one type of distribution with another

esp. using Normal




bounded / unbounded

https://commons.wikimedia.org/wiki/File:Binomial_Distribution.svg



28

## why is Normal normal?

central limit theorem

if you:

- average lots of things (or near linearly combine)
- around the same size (so none dominates)
- nearly independent
- and have finite variance

what?

then you get Normal distribution



29

## non-Normal – what can go wrong?

non-linearity – e.g. thresholds

+ve / -ve feedback

snowflakes and clouds

bi-modal exam marks

unbounded variance

the more you sample the bigger the variation

used to be rare e.g. wage/wealth distributions

… but now Power Law …

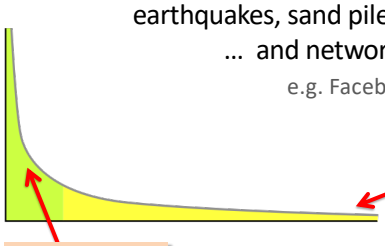


30

## power law – scale free

earthquakes, sand piles,
… and networks

e.g. Facebook connections



long tail
unbounded
variance

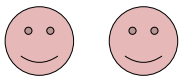
small number
of frequent
elements

power law data is NOT Normal
even when averaged

https://en.wikipedia.org/wiki/Power_law



31





32