Understanding Statistics for HCI and Related Disciplines

Part 2 – Doing It

if not p then what

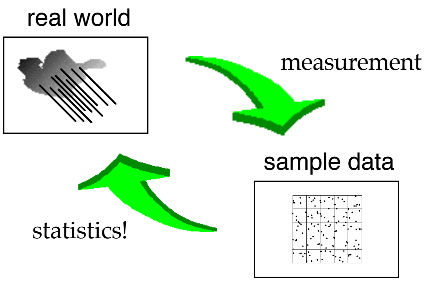
Alan Dix

http://alandix.com/statistics/

probing the unknown

quantified uncertainty

recall … the job of statistics



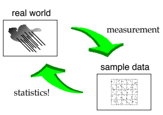
conditionals and likelihood

probability road will be busy
  vs probability busy given 4am on Sunday

probability die is a 6
  vs prob a 6 given it is 4 or bigger

conditional probability
  is probability given some knowledge

likelihood



likelihood is
  probability of measurement
  given unknown things in the world
  (conditional probability)

e.g. prob. of six heads given coin is fair = 1/64
  … but is it fair?

counterfactual

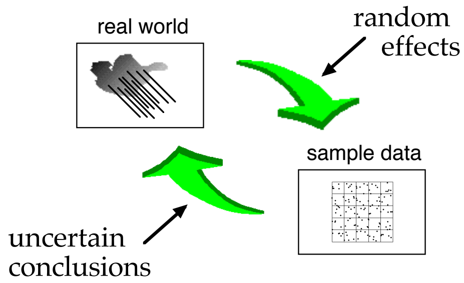
statistics turns this round

from likelihood of measurement
  probabilistic knowledge given the unknown world

to uncertain knowledge of the unknown world
  based on measurement

… but in various different ways

statistical reasoning



statistics

is never being able to say

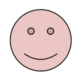
QED

types of statistics

- hypothesis testing (the dreaded p!)
  - robust but confusing
- confidence intervals
  - powerful but underused
- Bayesian stats
  - mathematically clean but fragile
  - issues of independence

same underlying methods

+ simulation methods for either



hypothesis testing

the ubiquitous p

significance test

hypotheses:

H_1 – what want to show

H_0 – null hypothesis (to disprove)

idea (when experiment/study successful)

if H_0 were true
then observed effect is very unlikely
therefore H_1 is (likely to be) true

5% significance level?

it says:

if H_0 were true
then probability observed effect happening by chance is less than 1 in 20 (5%)

so H_0 is unlikely to be true
and H_1 is likely to be true

does not say

✗ probability of H_0 is < 1 in 20

✗ probability of H_1 is > 0.95

✗ effect is large or important
i.e. significant in the real sense

all it says

✓ if H_0 were true ...

... effect is unlikely

(prob. < 1 in 20)

non-significant result

can NEVER reason:

✗ non-significant result ⇒ H_1 false ✗
things are the same

all you can say is:
H_1 is not statistically proven

a quick experiment

... but are they fair

calculations – six coins

given coin is fair:

- probability six heads = $1/2^6$ = 1/64
- probability six tails = $1/2^6$ = 1/64
- probability either = 2/64 ~ 3%

H_0 – coin is fair

H_1 – coin is not-fair

likelihood (HHHHHH or TTTTTT | H_0) < 5%

your experiment

toss 6 coins

record how many heads or tails

if HHHHHH or TTTTTTT

you can reject H_0 with $p < 5\%$

see how many times you do it before

you get 6 in a row

confidence intervals

bounds of uncertainty

proving equality

non-significant – not proved different

real difference may always be smaller than experimental error

⇒ can never prove equality

but can put bounds on inequality

confidence interval

bound on true value – same theory as p values

e.g. mean of data is 0.3

95% confidence interval is [–0.7,1.3]

says if the real value not in the range [–0.7,1.3]

probability of seeing observed data is less than 5%

counterfactuals

95% confidence interval is [–0.7,1.3]

does not say:
there is 95% probability that the real mean is in the range [–0.7,1.3]

it either is or it isn't!

all it says:
probability of seeing the observed data if real value outside the range is less than 5%

proven … what?

H_0: no difference (real mean is zero)

experimental result: mean is 0.3
significance test: n.s. at 5% – so what?
95% confidence interval: [–0.7,1.3]

? is 1.3 is an important difference

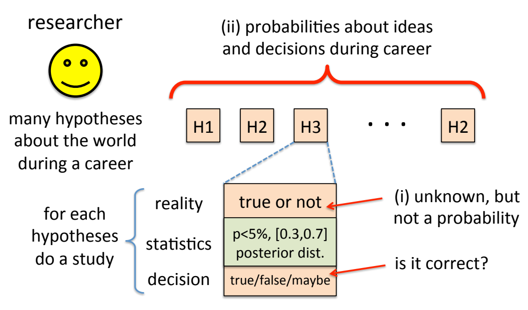
… and don't forget …

you still need to say
what test/distribution – e.g. Student's T
how many – degrees of freedom

it is still uncertain
the real value could be outside the interval

what you can say …

phenomena and statisticians



what you can say …

statement about decision during career not individual phenomena

if sig <5%
  and you act as if H1 as true
  then you are wrong at most 1 time in 20

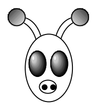
if you calculate 95% confidence interval
  and you assume true value is within CI
  then you are wrong at most 1 time in 20

Bayesian statistics

putting a number on it

traditional statistics

probability of having antennae:
  if Martian = 1
  if human = 0.001



hypotheses:
  H_0: no Martians in the High Street
  H_1: the Martians have landed

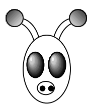
! you meet someone with antennae
  reject null hypothesis $p \leq 0.1\%$

call the Men in Black!

Bayesian thinking

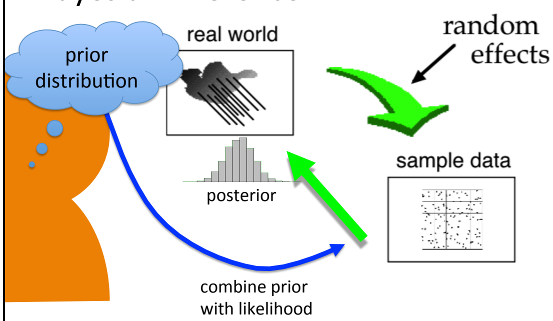
what you think before evidence

<u>prior</u> probability of meeting:
  a Martian = 0.000 001
  a human = 0.999 999
probability of having antennae:
  Martian = 1; human = 0.001



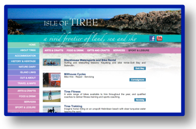
! you meet someone with antennae
<u>posterior</u> probability of being:
  a Martian ≈ 0.001
  a human ≈ 0.999

Bayesian inference



for adaptive websites

for adaptive websites



known visitors stats (*prior*):
surfers: 20%; others: 80%

imagine 100 visitors to website …

	surfers	others
all visitors:	20	80
click S&L:	15	10

first click to
sports & leisure:
surfers: 75%
others: 12.5%

so … if click S&L *posterior* probability
surfers: 60%; others: 40%

Bayes as a statistical method

new system B (*prior*):
A & B the same: 80%; B better: 20%

run a (v. small) study, all 4 users prefer B
prob. if A& B the same = 1/16
prob. If B better = 3/4 (say)

posterior probability
A & B the same: 25%; B better: 75%
odds ratio: 3:1

a different prior …

new system B (*prior*):
A & B the same: **20%**; B better: **80%**

same study, all 4 users prefer B
prob. if A & B the same = 1/16
prob. If B better = 3/4 (say)

posterior probability
A & B the same ~ **2%**; B better ~ **98%**
odds ratio: **48:1**

what are the prior and posterior?

sometimes

actual estimate of probability

e.g. patient with symptoms

more often

encoding **belief** as probability

phenomena is either true or not

Bayesian issues

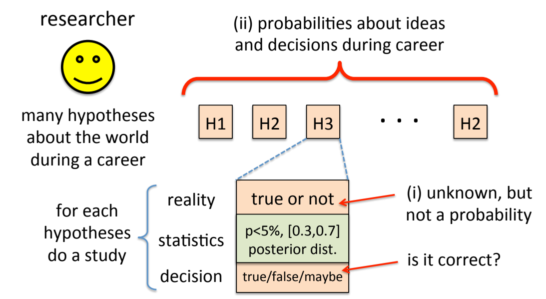
how do you get the prior?
- sometimes doesn't matter too much
- traditional stats rather like uniform prior
- beware confirmation bias!

handling multiple evidence
- can re-apply iteratively
- problems with interactions

internecine warfare
- traditionalists and Bayesians often fight ;)

what you can say … redux

phenomena and statisticians



what you can say …

statement about decision during career not individual phenomena

if prior is an accurate estimate of
the probability you are right
about hypothesis
of which you have similar belief
then the posterior is
the probability you are right this time

philosophical differences

probing the unknown

philosophical stance

? what do you know about the world

traditional statistics
- assume nothing!
- reason from unknowledge

Bayesian statistics
- 'prior' probabilities
- reason from guess-timates

uncertain and unknown

assumptions:
- traditional – NO knowledge of real value
- Bayesian – precise `probability' distribution

= measure of belief

in reality … sort of half know
- traditional – choice of acceptable p
- Bayesian – 'spread' distributions (uniform, Cauchy)

ideally should do sensitivity analysis … but no one does ☹

some dangers

if it can go wrong
then with 95% confidence …

inter-related factors

non-independent factors ← the same UI but 'not' consistent
e.g. fat and sugar

correlated features ← age and job seniority
e.g. weight, diet and exercise

avoid cherry picking

multiple tests ← survey with 40 questions

multiple stats ← including trying both traditional and Bayesian!

outliers

post-hoc hypotheses

…or worse

what to test for
and what to watch out for

what to test for?

10 coin tosses:
is it a fair coin?

what about … ?

Prob(THTHHHTTHH) = 2^{-10} ~ 0.001

Prob(HHHHHHHHHH) = 2^{-10} ~ 0.001

surely all heads is more extreme

the only thing 'the same or worse' than all heads is all tails:

but for
6 heads and 4 tails
nearly everything 'the same or worse'

easy for numeric data

we got 6 heads out of 10 tosses:

prob('the same or worse') assuming fair
= prob(anything but 5:5)
= 1 – prob(5 heads)
= 1 – 252/1024
~ 75%

another example

say 9 heads out of 10 tosses:

prob('the same or worse') assuming fair
= prob(9 heads) + prob(10 heads) + prob(9 tails) + prob(10 tails)
= 10/1024 + 1/1024 + 10/1024 + 1/1024
= 22/1024 ~ 2%

… and another

say 90 heads out of 100 tosses:

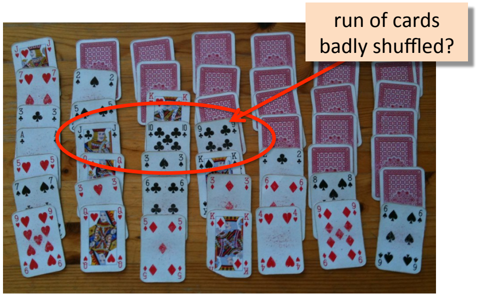
prob('the same or worse') assuming fair

= prob(nos heads >= 90)
+ prob(nos tails >= 90)

less than 1 in a million

but can be less obvious …

example: playing cards

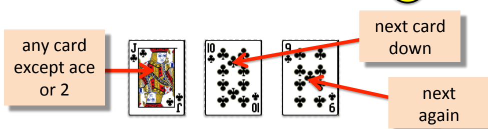


example: playing cards (2)

probability three cards make a decreasing run:

- first card needs to be 3 or bigger: 11/13
- second card exactly the next one down: 1/51
- third card exactly one down again: 1/50

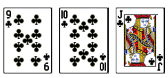
= 11/33150 approx 1 in 3000 – $p < 0.001$



example: playing cards (3)

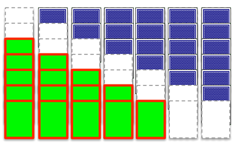
but equally surprised if …

- either direction
- 15 positions for the first card



= 15x2x11/33150

approx 1 in 100



and then …

how often do I play?

what to watch out for

I look at the data and notice

expert response time > novice resp. time

… hmm that looks interesting

I divide the data into experts and novices

… and do a t-test – Yay sig at 1% …

what else would have been equally interesting?

if 100 or even 10 then 1% doesn't say much!

so which is it?

to p or not to p
Bayes forward?

the statistical crisis?

mostly are well known problems with known solutions

maybe more people doing stats with less training?

some of the papers about it use flawed stats!

comparing

traditional statistics (p-test or confidence interval)
- quite conservative …
- … but only if done properly

Bayes
- many of the same problems as trad. stats. …
- … plus a few more!
- needs expertise

on balance (my advice!)

for most things stick with trad. stats.
… but always confidence intervals as well as p

for special things go for Bayes
- if you know prior
- decision making with costs
- internal algorithms

for both

do it properly

understand what you are doing
… and know what it means