**Part 2 – Doing It**

if not p then what

### Alan Dix

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

---

The job of statistics

**real world**

measurement

**sample data**

statistics!

---

**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

real world

random effects

sample data

uncertain conclusions

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

5% significance level?

you see it everywhere

sometimes significant at 5%
sometimes p<0.05 or p<0.01, p<0.001
sometimes p<5% or p<1%, p<0.1%
sometimes simply * or **, ***

what does it mean?

which is better?

5% (p ~ 0.05)
10% (p ~ 0.1)
1% (p ~ 0.01)
**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:

\[
\text{non-significant result} \implies H_1 \text{ false}  \quad \times
\]

all you can say is:

\[H_1\] is not statistically proven

---

**confidence intervals**

bounds of uncertainty

---

**proving equality**

non-significant – **not proved** different

real difference may always be smaller than experimental error

\[\implies \text{can never prove equality}\]

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?

H₀: 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

researcher

many hypotheses
about the world
during a career

for each
hypotheses
do a study

true or not

p<5%, [0.3, 0.7]

posterior dist.

true/false/maybe

(i) unknown, but
not a probability

is it correct?

(ii) probabilities about ideas
and decisions during career

H1 H2 H3 .. H2

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:
H₀: no Martians in the High Street
H₁: the Martians have landed

! you meet someone with antennae
reject null hypothesis p ≤ 0.1%

call the Men in Black!

Bayesian think

what you think before evidence

prior probability of meeting:
Martian = 0.000 001
human = 0.999 999

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

! you meet someone with antennae
posterior probability of being:
Martian ≈ 0.001
human ≈ 0.999
Bayesian inference

combine prior with likelihood

real world

random effects

sample data

prior distribution

posterior

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 ;)

33

Statistics for HCI – CHI 2022 – Alan Dix

34

Statistics for HCI – CHI 2022 – Alan Dix

35

Statistics for HCI – CHI 2022 – Alan Dix

36

Statistics for HCI – CHI 2022 – Alan Dix
what you can say ... redux

phenomena and statisticians

researcher

many hypotheses about the world during a career for each hypotheses do a study

(i) probabilities about ideas and decisions during career

H1 H2 H3 . . . H2

true or not
p<5%, [0.3,0.7] posterior dist.
true/false/maybe (ii) unknown, but not a probability

is it correct?

what you can say

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

statement about decision during career not individual phenomena
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

in reality ... sort of half know
  traditional  -- choice of acceptable p
  Bayesian     -- ‘spread’ distributions (uniform, Cauchy)

= measure of belief

ideally should do sensitivity analysis ... but no one does 😞
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 = $2/64 \approx 3\%$

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

see how many times you do it before
you get 6 in a row

the file drawer effect

you can only publish positive results
– the non-sig results go in the file drawer!

solutions
pre-registration – say what and how
reviewing method before results
some dangers

if it can go wrong then with 95% confidence ...

inter-related factors

- non-independent factors
  - e.g. fat and sugar
- correlated features
  - e.g. weight, diet and exercise

avoid cherry picking

- the same UI but ‘not’ consistent
- age and job seniority

- multiple tests
- multiple stats
- outliers
- post-hoc hypotheses

- survey with 40 questions
- p/b-hacking
- including trying both traditional and Bayesian!
seduced by numbers

dichotomous reasoning
  – 5% sig. is not true/false
  – levels of evidence

significance level is not probability
Bayesian posterior is not probability

why fixed values?

5%, 1%, 0.1% ?