The airport fallacy

Justin Smith's response ("An embarrassment of riches", page 46, October 2017) to our article "What are the odds!? The 'airport fallacy' and statistical inference" (August 2017) defends frequentist inference, stating that: "I simply fail to see how frequentism is a fallacy." Strangely, Smith did not actually address any of the arguments we made.

He first describes a simplistic coin-flipping experiment where there are little or no model selection issues that are typical of real data. But our central point was that data-dependent choices made during the model-building process for real, complex data afford a legion of opportunities for overfitting and, consequently, flawed inference. Moreover, without strict adherence to a pre-specified experimental protocol and statistical analysis plan – which would suffocate the exploratory nature of most scientific research – there isn't even a well-defined probability space on which to make such inferences. This is the airport fallacy. Coin flipping does not usefully represent this situation.

Smith next cites an example of 20 historical estimates of the solar deflection of light. He states that while individual experiments may have flaws, the estimates are consistent with general relativity, and confidence intervals are shrinking. Similar examples abound in the physics literature (e.g. the "Review of Particle Physics").[1] In his classic 1972 paper, "Enduring values", W. J. Youden discussed laboratory measurement examples in which different laboratories' estimates often disagreed by more than their reported error bars.[2] He interpreted this as evidence of unmeasured systematic error. This reinforces our point – statistical inference characterises uncertainty based on models of putative error sources, but such models are inherently oblivious to extra experimental systematic errors. This again implies that the greatest source of uncertainty stems from model choice: error bars obtained conditional on a particular chosen model understate this true uncertainty.

Smith next cites Deborah Mayo's work on severe testing. But Mayo defines severity using probability, thus requiring a model, so model uncertainty and overfitting issues still apply. Finally, Smith cites the argument that "likelihood swamps the prior" so that frequentist and Bayesian approaches usually tend to give similar results. Yet again this assumes reliable probability models and likelihood functions, ignoring the model choice and uncertainty issues that we argued are the heart of the problem.

Bert Gunter, Pleasant Hill, CA,
and Christopher Tong, Sparks, NV

References

1. C. Patrignani *et al*. (Particle Data Group) (2016) Review of Particle Physics. *Chinese Physics C*, **40**, 100001. See 2017 update at pdg.lbl.gov.

2. Youden, W. J. (1972) Enduring values. *Technometrics*, **14**, 1–11.

In their August 2017 article, Gunter and Tong question the widespread use of frequentist statistical inference. They identify just one place where the methodological problems that concern them may be avoided by rigorous design and pre-specification, namely in the conduct of pivotal clinical trials in medicine. They believe that the difficulties and resources required to protect reliable inferences are far too expensive and time-consuming for other areas of investigation. I write to contest this point.

I was heavily involved in the development of international guidelines for the design, conduct and analysis of clinical trials carried out for the purpose of licensing new medicines. These were the global ICH E9 guideline[1] and its European predecessor.[2] The inference problems that Gunter and Tong highlight were at the forefront of our minds in drafting these guidelines so that clinical trials worldwide would lead to reliable regulatory decisions.

In ICH E9 an important distinction is made between confirmatory and exploratory trials. As the name implies, a confirmatory trial is carried out to test a well-developed hypothesis generated by, and relying on, earlier work. This earlier work should provide all necessary information to aid the design of the confirmatory study and to allow the pre-specification of all important aspects of the conduct of the trial and its analysis. Some of the earlier work is likely to have been carried out in earlier exploratory trials, less rigorously designed and analysed.

It is hard to see why such an approach is not possible in most areas of scientific work covered by standard texts on the design and analysis of experiments. Gunter and Tong point to the time and cost involved. However, the cost of clinical trials is heavily influenced by the need to carry them out in a large number of patients, often in an international multi-centre setting, requiring major resources to recruit and manage the patients, monitor the study and collect data uniformly and reliably. Difficulties of this nature do not arise, for example, in studies carried out in laboratories, in many agricultural experiments, or in studies of manufacturing facilities. In addition, time invested at the time of design is always amply rewarded by savings of time in gaining acceptance of results, in implementing conclusions and in avoiding repetition of experiments.

John A. Lewis
Rothbury, Northumberland

References

1. Lewis, J. A. (1999) Statistical principles for clinical trials (ICH E9): An introductory note on an international guideline. *Statistics in Medicine*, **18**, 1903–1904.

2. Lewis, J. A., Jones, D. R. and Röhmel, J. (1995) Biostatistical methodology in clinical trials – a European guideline. *Statistics in Medicine*, **14**, 1655–1682.

On sample size

There are two points that one might add to Deirdre Toher's otherwise excellent and very thorough article ("Help! Is my sample big enough?", August 2017).

She points out that grinding up all the 40 flowers in the treatment

We welcome comments from readers – please email **significance@rss.org.uk**. Contributions should be no more than 300 words in length and clearly marked "for publication". Published responses may be edited to fit.

(and control) sample reduces variability and destroys information, but overlooks the more important defect that the three samples taken from the homogenised mass are not independent, so that even the Mann–Whitney test will not be appropriate as it assumes independence of observations. Better would be to take three non-overlapping subsets of the 40 flowers with 13 or 14 flowers in each, and homogenise these separately, to get three independent measurements in each group, which could then legitimately be compared.

Second, it could be questioned whether Mann–Whitney is really necessary. Taking measurements based on a homogenised sample from, say, 13 of the flowers amounts to taking the average of 13 random variables. The central limit theorem suggests that the mean of sufficiently many identically distributed random variables will be close to a normal distribution.

"Sufficiently many" is often taken to be around 30, and 13 or 14 might be rather small to ensure that using a parametric test is appropriate. Nevertheless, the fact that the distribution of a set of raw measurements may well be non-normal does not rule out the use of parametric tests if the data on which they are based are derived from averaging many of the individual measurements. Readers should not necessarily be put off using parametric statistics in such cases. As always, there is no fixed rule, but each instance needs to be judged on its own merits.

Dr Rory Allen
Goldsmiths, University of London

eriksvoboda/Bigstock.com

Wiley Prize Crossword: *TASS* by Goujeers

1		2		3	■	■	4	5		6		7		8
	■		■		■	9	■		■		■		■	
10									■	11				
	■		■		■		■		■		■		■	
12					■	13								
	■		■	■	■		■		■		■		■	
14				15			■	16				■	■	■
	■		■		■	■	■	■	■		■	17	■	18
■	■	■	19			20	■	21						
22	■	23	■		■		■		■	■	■		■	
24									■	25				
	■		■		■		■		■		■		■	
26					■	27								
	■		■		■		■		■		■		■	
28								■	■	29				

Send your solution to: *Significance* Crossword Competition, Royal Statistical Society, 12 Errol Street, London, EC1Y 8LX or scan it and email to significance@rss.org.uk. The competition is sponsored by Wiley (wiley.com/statistics), who will give the winner £100 or $150 to spend on Wiley books. Closing date: 12 January 2018. The winner will be chosen randomly from the correct entries, and the correct solution published in a future issue. Photocopies are acceptable.

Numbers in clues refer to members of a particular set, and may not be further defined.

Across

1, 12 Sounding rough and 16, there are many stories here (5,5)
4 1 soldier spies man outside (8)
10 Frees those with 'ounds getting het up inside (9)
11 21 turned for some to hear (5)
12 See 1
13 She inspired rock band with portent for us in muse (9)
14 Honest self-righteous person bathing in milk (7)
16 Part of graph held in both hands in book (4)
19 Sweet 0 (4)
21 Broken neck hit in mess (7)
24 Otter came transformed: it has eight feet (9)
25 15 was, on reflection (5)
26 Memory expert goes into space (5)
27 Are reflected hills following river upstream in mappings (9)
28 8's concentration of power (8)
29 That is missing from dogs outer layers (5)

Down

1 Ring missing ring? Tofu missing ring? Label is simple (8)
2 Drawn out horribly inappropriate (8)
3 Safe to echo (5)
5 Discharge sailor with crack (7)
6 With our thing elaborately embroidered (9)
7 State one's gods (6)
8 Prompted nude, good but missing a pair, to get dressed (6)
9 9 woman's college (6)
15 Changing room's managed to employ attendant (9)
17 Badges goat's cheese, leaving last bit and adding pounds regularly (8)
18 Study compulsive girl's chromosome first (8)
20 Frees Latvian holding old coin (4,3)
21 Core officer heard (6)
22 Solver may be 6 (6)
23 One looking at 17 with queen (6)
25 End of game 13 seen in side at Headingley (5)

Solution to October issue's crossword:
***Plain Crossword* by Sam Buttrey**

Across: 1 anag; **9** RATE F in GULLY; **10** O in MON; **11** hidden; **12** anag; **14** A VAN in HA; **16** anag; **18** anag; **20** hidden; **21** BAND AGED; **23** ALE + anag; **25** anag; **26** anag; **27** G (grand) in ROMAN, all in CONMAN.

Down: 2 EARL Y; **3** anag; **4** anag; **5** AT + ANT in LA; **6** (mon)KEY; **7** TAMP A; **8** hidden; **13** anag; **15** AVER AGES; **17** MOTHE(r) A TEN; **19** MAD ON NA; **22** hidden; **24** P ARK A; **26** hidden.

Winner: Albert Madansky, Chicago, IL

T	E	N	N	I	S	R	A	C	K	E	T	■	■	■
■	A	■	O	■	U	■	T	■	E	■	A	■	G	■
G	R	A	T	E	F	U	L	L	Y	■	M	O	O	N
■	L	■	R	■	F	■	A	■	■	■	P	■	V	■
Z	Y	D	E	C	O	■	N	A	Y	S	A	Y	E	R
■	■	■	D	■	C	■	T	■	U	■	■	■	R	■
H	A	V	A	N	A	■	A	R	G	U	M	E	N	T
■	V	■	M	■	T	■	■	■	O	■	O	■	E	■
S	E	L	E	N	I	U	M	■	S	A	T	U	R	N
■	R	■	■	■	N	■	A	■	L	■	H	■	■	■
B	A	N	D	A	G	E	D	■	A	L	E	P	P	O
■	G	■	O	■	■	■	O	■	V	■	A	■	A	■
G	E	A	R	■	B	A	N	N	I	S	T	E	R	S
■	S	■	I	■	O	■	N	■	A	■	E	■	K	■
■	■	■	C	R	O	M	A	G	N	O	N	M	A	N