A WAY OF TEACHING SIGNIFICANCE TESTING

J.H. Durran
Winchester College, U.K.

Nowadays increasingly many people are admitting and increasingly many people are claiming to be Bayesians. We have heard a lot already at this conference about Bayesian statistics. My subjective probability that I am a Bayesian is high. (Does that make me a Bayesian?)

We are not here, however, to talk about statistics but to discuss the teaching of statistics. To take what I believe is a parallel: we do not live in the inter-war years but I think it is important to teach people about the beliefs and values of that time. Our pupils are going to read many texts by non-Bayesians and they need to learn how to interpret the various terms involved. The concepts and methods to which those terms refer are rooted in common-sense. This emerges, I believe, when they are exposed (rather than taught) in the way I adopt and that I want to share with you. I am asking you to participate in a speeded-up version of what would take several sessions with pupils.

1. Can you catch me cheating, calling heads or tails spinning a coin? (Throw coin 30 or so times.) Was I cheating? How did you decide? I suggest the following:

   The "rest state" is No Cheating. The hypothesis that nothing unusual is going on is called The Null Hypothesis (NH). It can only be a hypothesis. (Do not tell whether cheating or not. Realism – if we could know, then statistical enquiry would be pointless.)

   The Alternative Hypothesis is that I was cheating. Consider some crucial event (e.g., wild imbalance or long run one way), some event whose occurrence would, if there was no cheating be improbable. If that events occurs then claim I was cheating, that is: choose to "reject" the NH; you cannot prove that I was cheating. Events that are probable under the NH are not indicative of anything useful. (Discuss the decisions of the class.)

   Calculate pr (test event occurs when NH is true). Adjust event to make the probability "small enough" (e.g., 0.5 or 0.1). This probability is the significance level of the test: low probability for high significance. We cannot calculate probabilities on basis of alternative hypothesis; it is too vague. (Clue for later :pr (six of a sort in a row) = $1/2^5$, not $1/2^6$.)

2. This time if I cheat, it will be towards Heads. Can you catch me? (Throw coin about 30 times. Discuss decisions. Most of class alter their test events to ignore long runs of tails and to look for shorter runs of

heads, and they do this before any theoretical discussion. They have taught themselves.)

This time the alternative hypothesis is directional "pr(head)>0.5. Using histogram, discuss "1-tail" and "2-tail" on basis that test event is imbalance of proportions.

3. This time you are fined 10¢ if you claim that I am cheating when actually I am not. Compare a manufacturer claiming that his medicines have "cheated" Nature and broughr benefits when actually they have not. (Throw coin about 30 times. Collect fines where appropriate. In discussion note that only more improbable events had triggered the accusation of cheating. The class teaches itself that that is the sensible strategy.)

   You want to avoid claiming cheating where there is not cheating; you choose pr(reject NH|NH is true) to be small; e.g., test might be 7 in a row rather than 6 in a row. (Do not get bogged down in actual probability values; it is their changing relative values that are the essence of the problem.)

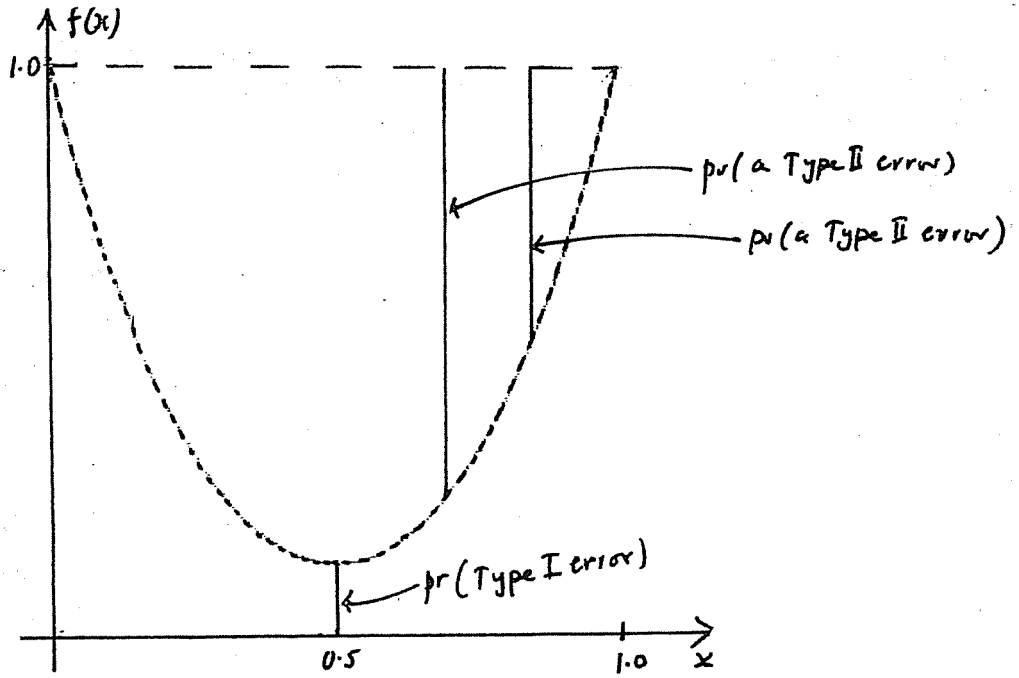
4. This time you are fined if you miss cheating. Compare a purchaser not noticing that he is allowing "cheating" by a manufacturer. (Throw coin about 30 times. Collect fines where appropriate. In discussion note that even only mildly improbable events trigger an accusation of cheating. Again class teach themselves.)

   You want to avoid missing cheating when there is cheating; you choose pr(accept NH|NH is false) to be small; e.g., test event might be 5 in a row rather than 6 in a row.

5. To reject NH when NH is true is Type I Error, see paragraph 3. To accept NH when NH is false is Type II Error, see paragraph 4.

   Conflict:pr(accept NH|NH is false) is small, which gives pr(reject NH|NH is false) is large, which gives pr(reject NH|NH is true) is large-ish, by continuity. See diagram:

f(x)=pr(reject NH|pr(head)=x).



Test needs to be well chosen. It is not enough to have significance high, that only means pr(Type I error) is small and may mean that pr(Type II error) is too high. It is often valuable to avoid both errors as far as possible. We go on to compare two simple tests using simulated throws already carried out and recorded.

6. Simulation of throws of coin via computer-generated random numbers

Null hypothesis: pr(even)=0.5; alternate hypothesis: pr(even)≠0.5;

Test A:
reject NH if first 4 results of a run are the same;

Test B:
reject NH if 6 or more of first 7 results of a run are the same;

Let pr(even)=x,

Test A:
$pr(\text{reject NH}|pr(\text{even})=x)=x^4+(1-x)^4=f_A(x)$

Test B:
$pr(\text{reject NH}|pr(\text{even})=x)=x^7+7x^6(1-x)+7x(1-x)^6+(1-x)^7=f_B(x)$

Significance level=pr(Type I error)=pr(reject NH|NH is true)

Significance level=pr(Type I error)=pr(reject NH | x=.05)=f(0.5)

Now $f_A(0.5)=f_\beta(.05)=0.125$, so for both tests significance level [=pr(Type I error)]=0.125.

We show results of 54 independent runs of 7 trials with pr(even)=0.5, i.e., with NH true. We use a/b to indicate Type I/II error using Test A. In both cases observed proportion of Type I error = 7/54 = 0.130.

1121 221	2111 211	2221 122	1112 111 b	1111 211 a b
2111 212	2112 111	2212 111	1122 212	2211 221
1121 121	1212 112	1122 111	2222 111 a	2212 111
2212 122	2112 221	2121 221	2111 112	2112 112
2211 212	2222 221 a b	1212 221	1222 212	1221 112
2222 212 a b	1212 112	1112 112	1111 222 a	1221 122
2221 112	2211 211	2211 212	1221 211	1211 111 b
1212 212	1122 221	2222 122 a b	2111 212	2111 222
2212 221	1212 211	1211 121	2111 211	1122 222
2122 221	1111 211 a b	1211 121	1122 111	2211 122
2221 122	2212 212	1122 122	1112 211	

7. Suppose that, unknown to us, pr(even)=0.12≠0.5, so that NH is false:
pr(this Type II error)=pr(accept NH | NH is thus false)
pr(this Type II error)=pr(accept NH | x=0.12)=1-pr(reject NH | x=0.12)

Now $f_A(0.12)=0.600$; $f_\beta(0.12)=0.799$;

so for Test A, pr(this Type II error)=0.40
for Test B, pr(this Type II error)=0.20

We show 54 independent runs of 7 trials with pr(even)=0.12, i.e., with NH false. We use c/d to indicate Type I/II error using Test A.

For Test A observed proportion of Type II error = 21/54 = 0.39,
For Test B observed proportion of Type II error = 10/54 = 0.19.

2222 212	1222 221 c d	2122 222 c	2222 222	2212 222 c
2222 222	2212 222 c	2222 222	2222 222	2222 222
2222 212	2121 222 c d	2222 222	2212 222 c	2222 212
2222 222	2222 222	2222 212	2222 222	1222 222 c
2222 222	2222 222	2221 222 c	2222 222	2122 222 c
2222 222	2222 122	2221 212 cd	2122 122 cd	2222 222
1222 222 c	2122 221 c d	2222 222	1222 222 c	2222 222
2222 222	2221 122 c d	2222 121 d	1222 222 c	2222 222
2222 122	2222 222	1221 222 c d	2221 122 c d	2222 222
2222 222	1222 222 c	2222 222	2212 212 c d	2222 122
2221 222 c	2222 222	2222 122	2222 222	

These matters are summarized on the graph which follows: