A Bayesish perspective on multiple choice score data

Bayesian Statistics While Standing on One Leg

A screening test for a rare disease is known to be 99.9% accurate.
You test positive. What is the probability that you have the disease?

Case 1: Low-Risk Population
Incidence is 1 in 1,000,000
So out of 1,000,000 people:

- 1 person is sick
- 999,999 are not

Of those who test positive:

- 99.9% are well!
- Only 0.1% are sick

	Sick	Not Sick
Test +	1	1,000
Test -	0	998,999

Case 2: High-Risk Population
Incidence is 1 in 1,000
So out of 1,000,000 people:

- 1,000 people are sick
- 999,000 are not

Of those who test positive:

- 50% are well!
- 50% are sick

	Sick	Not Sick
Test +	999	999
Test -	1	998,001

A simple mathematical model of 4-choice multiple choice testing situations

- A known proportion of students, *p,* select the correct response.
- Some unknown proportion of students, *k,* **know** the correct answer.
- Everyone else guesses; 25% of them guess right.
- So in every case, $k \leq p$.
- We can say, $p = k + \frac{1}{4}(1-k)$ (just an estimate!)
- We can solve for k (again, just an estimate!): $k = (4/3)p - \frac{1}{3}$
- Intuitively, if 1,000 students are asked a 4-choice question and they all choose the right answer, no one guessed. Conversely, if 25% choose the right answer, everyone guessed.

If 40% of students chose the right answer to an item (*p=0.40*), half of them guessed
If 60% of students chose the right answer to an item (*p=0.60*), only 22% of them guessed

But we can do better than this.
We don't just have student responses to individual items.
We also have overall scores, and we can segment the overall population on that basis.
Suppose that the p-value for an item is 0.25 for the bottom quartile, and 0.90 for the top quartile.

- Of bottom quartile students *who got this item correct*, all guessed! **None** of them actually knew the right answer.
- Of top quartile students who got this item correct, only 4% guessed! 96% of those who got it right did so because they knew the right answer.

“Bayesish”???

Not quite a Bayesian approach.
A student’s overall score isn’t exactly a “prior probability”.
But by treating it as such, we can take advantage of Bayes’ theorem

p	0.1	0.2	0.25	0.3	0.4	0.5	0.6	0.7	0.8	0.9
k	-0.20	-0.07	0.00	0.07	0.20	0.33	0.47	0.60	0.73	0.87
p-k			0.25	0.23	0.20	0.17	0.13	0.10	0.07	0.03
(p-k)/p			1.00	0.78	0.50	0.33	0.22	0.14	0.08	0.04
k/p			0.00	0.22	0.50	0.67	0.78	0.86	0.92	0.96

Completing the analogy

A screening test for proficiency in a skill is known to be 75% accurate.
A student tests positive. What is the probability that she is proficient?

Case 1: Low-Proficiency Population
Incidence is 1 in 10
So out of 1000 students:

- 100 are proficient
- 900 are not

Of those who test positive:

- 75% are not proficient
- 25% are proficient

	Prof.	Not Prof.
Test +	75 / 100*	225
Test -	25 /0*	675

Case 2: High-Proficiency Population
Incidence is 9 in 10
So out of 1,000 students:

- 900 are proficient
- 100 are not

Of those who test positive:
At least* 96.4% are proficient; no more than* 3.6% are not

	Prof.	Not Prof.
Test +	675 / 900*	25
Test -	225 / 0*	75

*In fact, “negative tests” are much more than 75% accurate. The starred numbers are what you get if you assume that the false negative rate is zero, meaning that proficient students never test negative.

Case study - A Beacon math grade 3 benchmark test

28 points possible: 18 MC, 2 2pt MCAo, 1 2-pt CR, 1 4-pt CR
n = 2,696
Segmented students by overall score, adding adding scores symmetrically as necessary to ensure at least 30 students per segment.

- e.g., 306 students scored 9 (out of 28), so students scoring 9 were segmented on their own
- Only 14 students scored 21. To get a large enough group, students scoring 21 were segmented with students scoring 20 to 22 (n=42)

Two different approaches were tried for students scoring below the expected score for random guessing (5.4).

- As described above (not shown here)
- all such students treated as one segment (n=272) (shown here)

Where the model estimates $k \leq 0$, I substituted $k = 0.002$

3.MD.A - 1 point from 1 MC item

Raw Score	Segment	p	k	P(K|1)
5 or less	0 to 5	0.21	-0.05	0.01
6	6	0.33	0.10	0.31
7	7	0.46	0.27	0.60
8	8	0.57	0.43	0.75
9	9	0.64	0.52	0.81
10	10	0.73	0.64	0.88
11	11	0.85	0.80	0.94
12	12	0.88	0.83	0.95
13	13	0.93	0.90	0.97
14	14	0.94	0.92	0.98
15	15	0.96	0.94	0.99

If you are setting up an intervention for students who did not know the content on this question, and you want to include anyone who is less than 80% likely to have actually known the correct answer, then you should include not only all students who got the question wrong, but also ALL students who got it right, but also had an overall score of less than 9.

On this question, random guessing was not a material factor for students with an overall score of 12 out of 28 (43%) or higher

3.MD.A - continued

Raw Score	Segment	p	k	P(K|1)
16	16	0.95	0.94	0.98
17	17	1.00	1.00	1.00
18	18	1.00	1.00	1.00
19	18 to 20	0.99	0.98	1.00
20	19 to 21	0.98	0.98	0.99
21	20 to 22	1.00	1.00	1.00
22	20 to 24	1.00	1.00	1.00
23+	20 and up	1.00	1.00	1.00

3.OA.A - 2 points from 2 MC items

Raw Score	Segment	P2	K2	P(K|2)
5 or less	0 to 5	0.11	.05	.46
6	6	0.22	0.16	0.76
7	7	0.31	0.26	0.85
8	8	0.41	0.37	0.90
9	9	0.46	0.43	0.92
10	10	0.55	0.52	0.94
11	11	0.56	0.54	0.95
12	12	0.65	0.62	0.96
13	13	0.67	0.65	0.97
14	14	0.74	0.73	0.98
15	15	0.70	0.68	0.97

In this case, we assume that anyone who scores below 2 on these 2 items is less than proficient in the cluster. What we want to know is, of those who scored 2 (p2), what is the probability that they knew the answers versus just having gotten lucky? (¼ x ¼ = 1/16 = .0625)

At the very bottom of the score distribution, more than half of the 2s are attributable to guessing. But for the rest of the score distribution, guessing plays only a small role.

3.OA.A - continued

Raw Score	Segment	p2	k2	P(K|2)
16	16	0.83	0.81	0.99
17	17	0.68	0.66	0.97
18	18	0.68	0.66	0.97
19	18 to 20	0.80	0.78	0.98
20	19 to 21	0.88	0.87	0.99
21	20 to 22	0.91	0.90	0.99
22	20 to 24	0.91	0.91	0.99
23+	20 and up	0.92	0.91	0.99