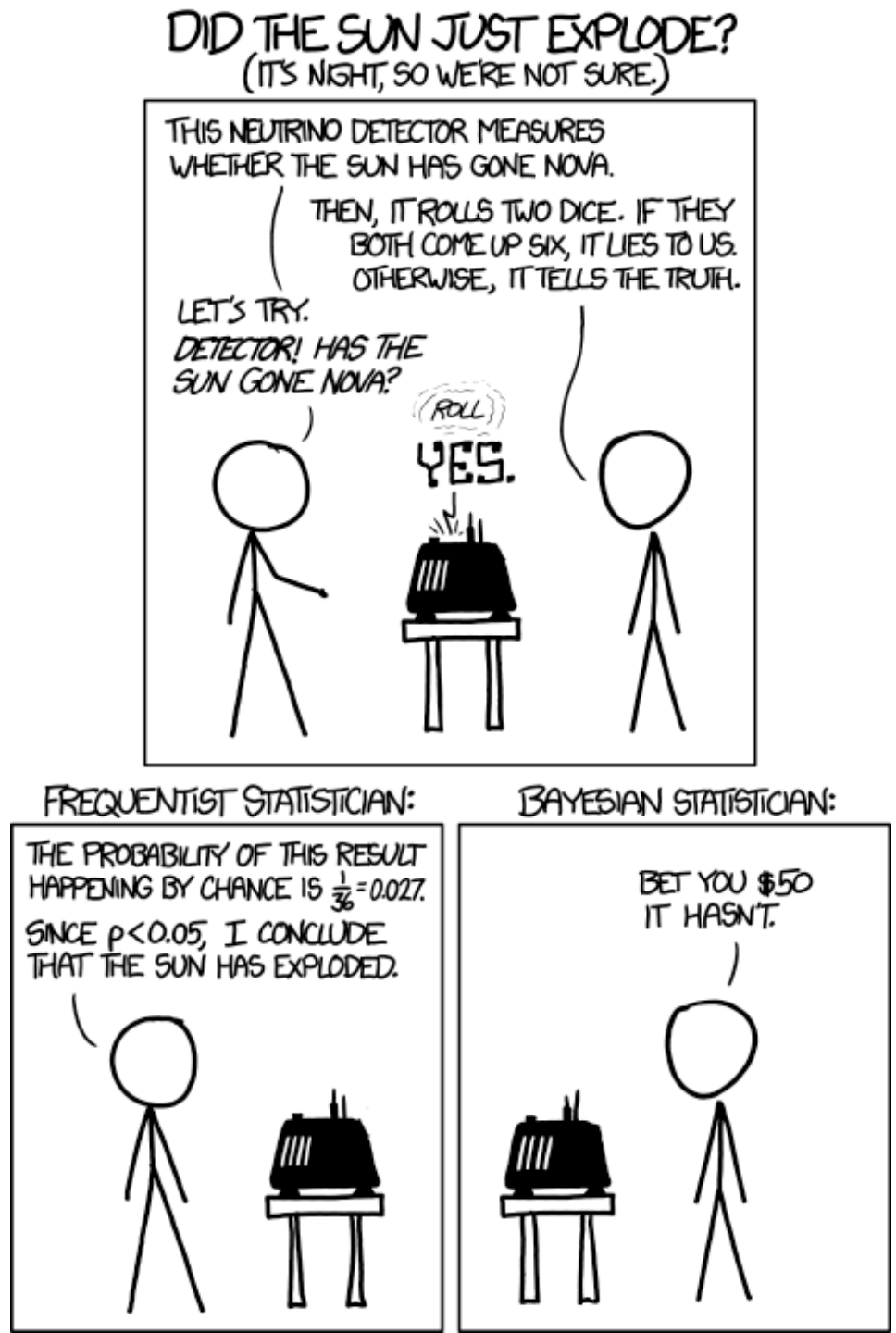
Pre-class assignment

1) In Bayes' Theorem, what is the utility of including the prior and evidence probabilities? (In other words, why is it useful to include them?)

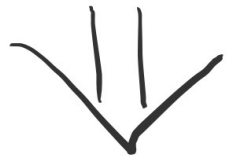
2) What is the difference between the frequentist and Bayesian approaches toward statistical inference, and how does it manifest in the way these approaches are used?



https://xkcd.com/1132/

Bayes' theorem:

$$P(x|y) = \frac{P(y|x) \times P(x)}{P(y)}$$



$$\left[\begin{array}{l}\text{probability of } x \\ \text{given } y\end{array}\right] = \frac{\left[\text{prob. of } y \text{ given } x\right] \times \left[\begin{array}{c}\text{prob. of} \\ x\end{array}\right]}{\left[\text{prob of } y\right]}$$

or

$$[\text{posterior}] = \frac{[\text{Likelihood}] \times [\text{prior}]}{[\text{evidence}]}$$

In terms of a model (w/ params Θ) and data (d):

$$p(\Theta|d) = \frac{p(d|\Theta)\,p(\Theta)}{p(d)}$$

or

[probability that model parameters Θ describe this data d] = [probability of data given model parameters] × [prob of model parameters] / [probability of data]

[probability that model parameters Θ describe this data d] — posterior

[probability of data given model parameters] — likelihood

[prob of model parameters] — prior

[probability of data] — Evidence

Question You receive a test for a certain disease. 2% of people in your demographic have the disease, and the test is 90% reliable. If the test is positive, how likely are you to actually have the disease?

Bayesian approach:

$$P(\text{disease given positive test}) = \frac{P(\text{test given data}) \times P(\text{data})}{P(\text{test being positive})}$$

$P(\text{disease given positive test}) = P(D|T)$

$P(\text{test given data}) = P(T|D)$

$P(\text{data}) = P(D)$

$P(\text{test being positive}) = P(T)$

$$P(D|T) = \frac{P(T|D)\,P(D)}{P(T)}$$

$P(T|D) = 0.9$ = accuracy of test

$P(D) = 0.02$ = chance of a person in your demo having this disease

$P(+)$ = (prob. of true positive) + (prob. of false positive)

$= P(T|D)P(D) + P(T|\text{not } D)P(\text{not } D)$

$= 0.9 \times 0.02 + (1-0.9) \times (1-0.02)$

$P(D|T) \simeq 0.1552 \sim 15.5\%$

Followup: how does the probability of you actually having the disease change as a function of test reliability? (Try reliability values of 0.8, 0.95, 0.99, 0.999)

reliability	P(DIT)
0.75	0.0577
0.8	0.0754
0.9	0.1552
0.95	0.2794
0.99	0.6689
0.999	0.9532
0.9999	0.9951
0.99999	0.999

Context: COVID-19 antigen tests

Are ≈ 80% reliable, and false positive and negative results may differ.

COVID-19 molecular tests are closer to 99-99.9% reliable.

Question: If I have photometric observations of a galaxy that *might* be a high-redshift galaxy but *could* be other things (e.g., a closer, red galaxy), how do I assess the Bayesian probability of the high-z hypothesis?

Some hints:

1) What do you know about galaxy evolution over time?

2) How do you include both true positives and false positives?

probability of high-z hypothesis

likelihood that high-z galaxy model matches data

prior based on galaxy distribution as f(redshift)

$$P(\Theta|d) = \frac{P(d|\Theta)\,P(\Theta)}{P(d)}$$

$P(d)$ = P(prob. of data match to model)

false positives

$$P(d) = \underbrace{P(\text{data fit to high-z model})\,P(\text{high z galaxies})}_{\text{true positives}} + P(\text{data fit to low-z model}) \times P(\text{low-z galaxies})$$

plugging in some (hand-wavy and unreliable) numbers:

$p(d \mid \text{high } z) = 0.7$ [high-redshift model matched to data]

$p(d \mid \text{low } z) = 0.4$ [low-redshift model matched to data]

note: these do not need to add up to 1. Why?

$p(\text{high } z \text{ gal}) = 0.1$

$p(\text{low } z \text{ gal}) = 0.9$

} probability representing underlying galaxy distribution (more galaxies at late time!)

$$p(\text{high } z \text{ galaxy} \mid \text{data}) = \frac{0.7 \times 0.1}{0.7 \times 0.1 + 0.4 \times 0.9}$$

$$= 0.16279 \rightarrow \approx 16.3\% \text{ galaxy is at high redshift!}$$

notes:

- underlying sample matters: more galaxies at low redshift are observable! (so even a low prob. of error for a given galaxy impacts the outcome significantly)

- model uniqueness/data fit matters a lot: this is why few-band photo-z is awful, many-band helps w/excluding false positives

photo-z PDF for a $z \sim 6$ blue galaxy w 5 HST bands might look a lot like this: