WHY BAYES?

> There is more in common than separates frequentist and Bayesian regression. The main question of both is how to translate fitted models into scientific arguments. This question is essentially unsolved.

Neither frequentist nor Bayesian regression automates scientific inference. Statistical models quantify known unknowns, but in the real world there are unknown unknowns, which cannot be built into the model.

Bayesian vs. frequentist comparison

1) philosophically, the real-world interpretation of mathematical probability is different: Frequentist probability is an objective feature of what happens in an expreimental system – Pr = long run relative frequency of events. Pr can only be meaningfully applied if there is a stable long-run frequency of successes/events. Hypotheses/parameter values do not have such a frequency. Bayesian probability quantifies uncertainty. It is purely a mental construct, which does not exist independently of a humming brain. For a Bayesian, the frequentist probability is a special case of bayesian probability, where knowledge can be described as frequencies. For a frequentist, Bayesian probability is a contradiction in terms, like a blue smell.

2) different goals: The frequentist goal is to reduce/summarise numerical data into a few statistics. Its product is point estimates + p values/CI-s. It prefers statistical procedures with certain *ad hoc* properties (**unbiasedness** – estimates are correct on average; **coverage** – CI-s cover the true parameter value with a given frequency; **conservatism** – weak data do not lead to strong inference, but they will still lead to unbiased inference with set coverage). The goal is an objective path from data to conclusions. Bayesian inference is applied probability theory, so it is theorethically the best way of combining uncertainty, and it need not to worry about any *ad hoc* test properties. Bayesian inference starts from uncertain knowledge and carries this uncertainty into conclusions. It merely converts unknown unknowns into known unknowns. Bayesian 90% CI-s tells that the model is 90% sure that true value lies inside the interval. Prior knowledge and new data have equal epistemic weight. **So, where the frequentist reduces data, bayesianist integrates data with prior knowledge. For the bayesian the goal is to model the process that generated the data, not the data as such.**

3) different modelling strategies – Bayesian strategy is to fit several models, study and discuss them comparatively. Its an iterative

process of model criticism & model improvement. Bayesian models are generative – you can first build a model using your prior knowledge, then generate fake data from the model, and only then fit the model on real data. You hope that by looking at all the models you fitted, you can test your hunches about the process that generated your data. data generating process –> model structure + data –> model fit –> model generated data & counterfactuals –> improved model –> n iterations –> real world conclusions.

4) different technical strenghts. Frequentism fixes long run error frequencies, essentially providing quality control for the experimental system (but not directly to individual experiments). Also, model comparison can usefully be frequentist. The Bayesian strength lies in multilevel/hierarhical/mixed effects models – easier to build, generally better results with less hassle. Bayesian modelling is generally more flexible, which is the major selling point for statisticians.

5) different impact on scientific conclusions. Frequentist stats curbs your enthusiasm by pointing out that p>0.05, and thus your exp system/sample size/effect size is insufficient for strong conclusions. A Bayesian model spits out a numerical strenght of belief, which you can accept as such, if you trust the model. You don't care about point estimate (which can be biased) but about the whole posterior, which is summarised by credible interval. Bayesianism is an integral part of scientific thinking. Frequentism distorts scientific thinking (its insistence of pre-deciding stat analysis before seeing the data, stopping rules, drawing decision lines for continuous outcomes).

6) different mathematical approach. In bayesian stats data is fixed (you are conditioning the model on your exact data) and parameter values are unknown (in stats-speak prameter values vary, until you fix them by fitting the model on data). you ask: What are the probabilities of parameter values, conditional on data & prior knowledge? In frequentist stats it is data that varies (you are looking at all possible data under a null hypothesis) and parameters are fixed as the null hypothesis. Because parameter values have no frequency, they cannot have probabilities, and they must be fixed not on data, but by fiat, as H_0. Thus the frequentist question is: what is the probability of data as extreme or more extreme than yours, conditional on the fixed parameter value (the null hypothesis)? So bayes: $P(H_1|data, prior)$; frequentist: $P(data|H_0)$. The only formal route from $P(data|parameter)$ to $P(parameter|data)$ is the Bayes theorem, which contains $P(parameter)$, i.e. the prior.

Bayesian regression differs in:

1) overfitting can be easily reduced by setting priors. You can easily study how much adding data to the model changes model predictions over priors alone. Priors represent your knowledge on the data generating mechanism independent of data. So its possible to derectly see, how much data adds to existing knowledge.

2) All Bayesian models are generative – you can generate new data from a fitted model and study, how these data are similar/different to real data. You can also re-fit the same model on data generated by the model to see, how coefs change. This diagnoses problems where model structure does not reflect data generating process.

3) Theres more flexibility in setting up models – you can easily do stuff that cannot be (easily) done in frequentist paradigm

4) Bayesian models tend to work better on poor data (small N, and/or y data distribution is far from normal and/or biased). Frequentist models are created for large-sample settings, they often break down when N is small. Bayesian models are better, because they use data datapoint-by-datapoint (and work as a matter of logic as well with N=1 than with N=1000), and because they use prior knowledge, which can alleviate the harm done by small and biased datasets.

5) multilevel modelling is much better in the Bayesian paradigm!

Probability theory is an axiomatic system.

Any real number that behaves according to following 4 rules is "probability".

1) probability $P(A) \geq 0$

2) the probabilities of mutually exclusive and exhautive hypotheses/events sum to 1. a.k.a. the sample space/hypothesis space omega sums to 1: $\Omega = 1$

3) for mutually exclusive independent events: $P(A \vee B) = P(A) + P(B)$

4) conditional probability is defined as $P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$.

- P(A) = 0 means tha A is FALSE (impossible)
- P(A) = 1 means that A is TRUE (certain)

note: A and B are mere placeholders for anything that behaves according to these rules. We can call them "event", "hypothesis", "data", "success", or whatever. $\vee$ is logical (non-exclusive) "or" and $\wedge$ is logical "and". If 6 seems non-trivial, as an exercise, substitute A with "it rains" and B with "its cloudy" and think it through.

Probability theory is a model for rational thought under uncertainty.

If probabilities take only two values, 0 and 1, then Pr theory reduces to propositional logic. This means that Pr theory is consistent with ordinary logic, and is a richer version of it. Propositional logic is monotonic – once you have reached a conclusion based on premises, new data cannot change that conclusion. Conversely, in probability theory uncertain evidence cannot lead to certain conclusions, making new data always relevant.

Pr theory is the best possible way of assigning plausibilities to theories, but it works only if (i) the hypothesis space is correctly specified (no relevant hypothesis is missing), (ii) all relevant information is used in calculations. In effect, Pr theory teaches how to be perfect, which does not necessarily mean that it teaches how to be good.

using pr theory in statistical inference

Bayes theorem is an elementary deduction from 6.

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(B)}$$

where P(B) is the normalization member, which guarantees that probabilities sum to one.

$$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + ...$$

for mutually exclusive and exhaustive A-s that exhaust the hypothesis space.

If we define A as "hypothesis" and B as "data" (lets say that the data consists of an Elvis sighting), and the hypothesis space consists of 2 hypotheses (say, Elvis is dead/Elvis is alive), then

$$P(alive \mid sighting) = \frac{P(alive)P(sighting \mid alive)}{P(alive)P(sighting \mid alive) + P(dead)P(sighting \mid dead)}$$

If hypothesis space has more members, then there will be more members under the line.

- $P(alive \mid sighting)$ - posterior Pr, or "probability that Elvis is alive, given the sighting"

- $P(sighting \mid alive)$ - likelihood, or "probabiity of the sighting, given that Elvis is alive", or "how likely it is to see these data, if the hypothesis is true". Likelihoods are numbers between 0 and 1, but they need not sum to one. Thus, likelihoods are NOT true probabilities.
- $P(alive)$ - prior Pr, or "how likely it is that Elvis is alive, disregarding the sighting". Every hypothesis in the hypothesis space has a prior Pr, and priors sum to 1.

The important part is to remember that the posterior is proportional to likelihood times prior.

Example (with real US data): We have an antibody test for COVID-19. If 1000 symptomatic covid patients are tested, on average 840 of them get a positive result (sensitivity of the test is 84%). If 1000 non-infected people are tested, 5 get a positive result (specificity of the test is 1 - 0.005 = 0.995, or 99.5%). We think that about 1% of the population has been exposed to the virus. Now, if a random person gets a positive result, what is the probability that this person is infected?

- hypothesis space has 2 members: H1 - infected, H2 - not infected
- priors: P(H1) = 0.01, P(H2) = 1 - P(H1) = 0.99
- likelihoods P(+ | H1) = 0.8, P(+ | H2) = 0.05

```
Pr_H1 = 0.01
Pr_H2 = 1 - Pr_H1
L_H1 = 0.84
L_H2 = 0.005
(Pr_infected = (Pr_H1 * L_H1)/(Pr_H1 * L_H1 + Pr_H2 * L_H2))
```

```
## [1] 0.6292135
```

There is a more intuitive solution. Imagine 1000 random persons, 10 of whom are infected and 990 are not. Now from the infected 10 x 0.84 = 8.4 test positive and from the uninfected 990 x 0.005 = 5 test positive. The probability of being infected, if getting a positive test result, is then 8.4/(8.4+5) = 0.63.

Ok, we now have a 14% probability that the person is infected, and corresponding 86% probability that she is not infected (that the test result is false-positive). This is why it may not be a good idea to test random individuals for stuff like covid, breast cancer or prostrate cancer, whose incidences are low and tests have less than perfect specificities.

What happens if we test with the intention of estimating true incidence rates in populations? Then we can take sensitivity/specificity into account and still get a good estimate. Soon we will do exactly this.

NYT 24.08.2020: Getting an antibody test to see if you had Covid-19 months ago is pointless.

Many tests are inaccurate, some look for the wrong antibodies and even the right antibodies fade away, said the Infectious Diseases Society of America, which issued the new guidelines.

Antibody testing generally should be used only for population surveys, not for diagnosing illness in individuals, the panel said. Even for that purpose, only tests that are correctly positive more than 96 percent of time and correctly negative at least 99.5 percent of the time should be used, according to the guidelines. Very few of the dozens of tests that the panel looked at met those standards. None can be done at home or immediately in doctors' offices, and the best are assays known as Elisa or chemiluminescence immunoassay.

With two exceptions, antibody tests should not be used to diagnose individual infections, the society said. When a patient has all of the symptoms of Covid-19, including X-ray evidence of pneumonia, but still comes up negative on repeated diagnostic PCR tests for the virus, an antibody test may be useful.

The tests can also be used for diagnosis when a doctor suspects a child has multisystem inflammatory syndrome, a rare but serious complication of Covid-19 in children. Because it is not known how long after the initial infection this inflammation begins, doctors should do both a PCR test and an antibody test, the guidelines said.

What if we want to estimate the true value of a continous parameter (for example the probability that a patient dies of COVID_19, or the mean length of hospitalization with COVID-19 in days)? In such cases the hypothesis space consists of infinite number of elements (real numbers from 0 to 1, and integers from 0 to Inf, respectively). This means that we must specify an infinite number of likelihoods and an infinite number of prior probabilities, before we can run the calculation! Luckily we can do this using continuous functions, which by definition specify infinite number of points. Then the posterior will be a continuous function as well.

what does Bayesian statistics do?

It synthesises existing knowledge with new knowledge (data, information). We always start with defined prior probabilities over the hypothesis space and then we update these probabilities according to new data. As the hypothesis space always sums to one, when the porobability of one hypothesis rises, some other hypotheses must become less likely. Therefore, doing science is really shifting the probabilities.

In practice, the output of a Bayesian model is a posterior distribution for each parameter in the model, which is then reduced for presentation purposes to a list of most likely parameter values accompanied with credible intervals (CI). A 90% CI means that the model says that the true parameter value lies inside this interval with 90% probability.

What is the meaning of bayesian probability

You have calculated that the posterior probability of it raining tomorrow is 0.7. What does this number mean? Think of it like this: you are contemplating buying a ticket that pays 1€ if it rains tomorrow (and nothing if it does not). Now, assuming that you are satisfied that you have managed to incorporate all your prior knowledge into the model, as a rational person you are should spend no more than 70 cents on this bet, which will then give you a 30-cent profit (3 to 7 odds). Should you go through a long life making such maximum bets, then, if you are a good modeller, you expect to come out about even on your deathbed.

When not to do bayes:

1) When you actually want to fix the frequency of type 1 errors. This may happen in quality control settings.

2) When you have plenty of data, and no useful prior information, then frequentist tests will give (nearly) identical results quicker. There are over ten thousand such tests, and if you find one that correponds well to your experiment & data, why not use it.

3) Bayes requires distributional models, both for priors and for likelihoods. The goal is to use the scientific information available to you as efficiently as possible – probability distributions for data and prior knowledge are there to represent this information in mathematical form. If you do not have the scientific knowledge to put into probability distributions, then consider bootstrap or (non)parametric frequentist tests.

4) Bayesian models run more slowly, and sometimes too slowly.

Textbooks et al.

From simplest/most elementary to more advanced texts:

- Gelman, Hill & Vehtari "Regression and Other Stories" (2020) – An up-to-date practical introduction/handbook to Bayesian applied regression with lots of code examples. This book is essential, while dry, reading. It does not cover multilevel models.
- Richard McElreath "Bayesian Rethinking", 2nd ed. (2020) – Gives a strong conceptual understanding of Bayesian inference, incl. the multilevel stuff. Also comes with practical code examples.
- John Kruschke "Doing Bayesian Data Analysis, 2nd ed." (2015) - more mathematical and less fun than McElreath. Covers the same ground, more or less.

  Web:

- `https://bookdown.org/ajkurz/Statistical_Rethinking_recoded/` - Statistical rethinking recoded into brms & ggplot2. A very useful companion to the McElreath book.
- Andrew Gelman blog - general discussion of stats in social science – lots of examples of bad statistics
- Stan user guides. – narrowly focused on Stan code. `https://mc-stan.org/users/documentation/`
- Paul Bürkner brms pages. – narrowly focused on brms code. `https://github.com/paul-buerkner/brms`