National Centre for Earth Observation
NATURAL ENVIRONMENT RESEARCH COUNCIL

NERC SCIENCE OF THE ENVIRONMENT

Same basic thoughts on Classical and Bayesian Inference

Peter Jan van Leeuwen

How do we process new data?

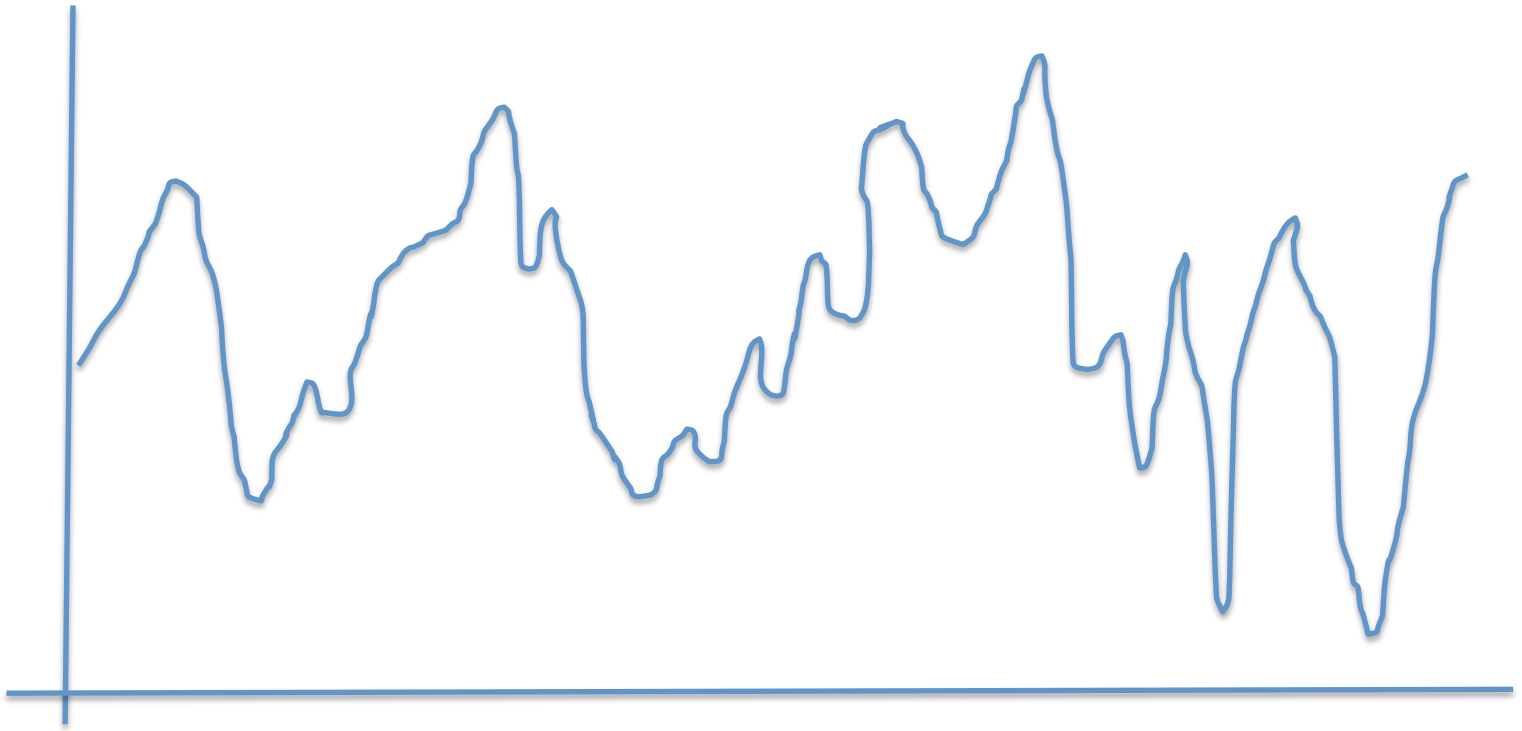
A process description

- Prior knowledge, from you model of the world, a cat
- Observation, the dog
- Posterior knowledge, improvement of the model, the dog that has eaten the cat

The human learning cycle

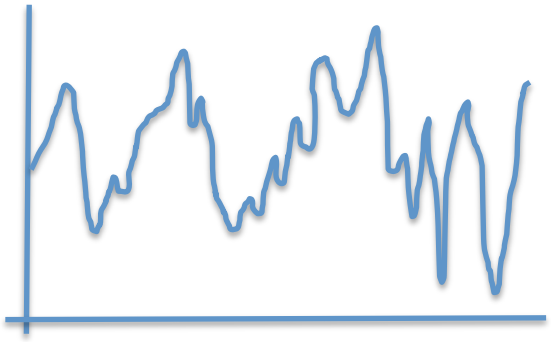
- We have a view of the world, the prior
- New observations come in, the likelihood
- We combine the two, and form the posterior
- The posterior is our new prior
- New observations come in, the likelihood
- We combine the two, and form the posterior
- Etc.

Example



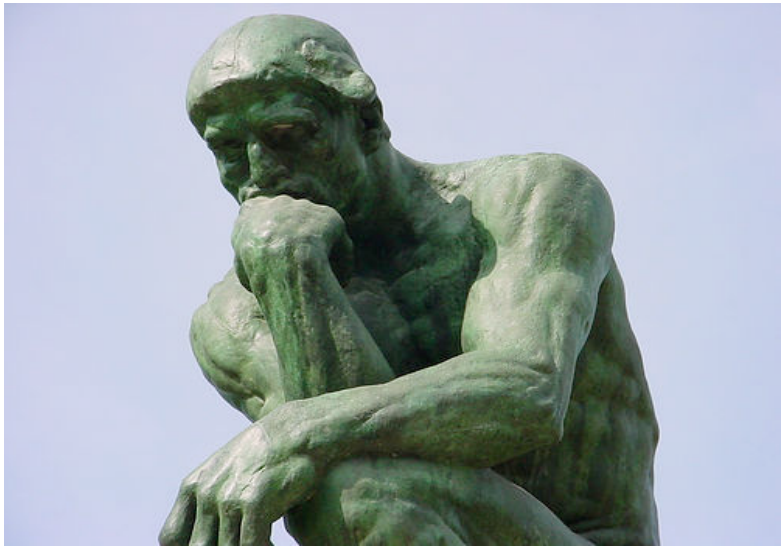
Is there a trend?

Is there a trend?



Trend ?

YES !



Questions…

- How certain is this result?
- How robust is it?
- How did the machine come to this answer?
- I have other indications that tell me there is no trend, how do I take these into account?
- Am I happy with a yes/no answer (am I a politician) ?

- How would I like the machine to come to an answer?

Science 101

- We want to infer something about the trend
- And we want to use the observations (or model results) as basis.
- Formally, we want to infer about the trend *given* the observations.
- We want to reflect that observations have uncertainties, so *our inference will also be uncertain*.
- One way to quantify this uncertainty is via a probability
- Doing that we find that we want to know
- *trend* means size of the trend

$$p\left(\, trend \;\middle|\; observations\right)$$

How do we calculate this probability?

Use Bayes Theorem:

$$p\left(trend \mid obs\right) = \frac{p(obs|trend)}{p(obs)} p(trend)$$

Likelihood

Prior

We need to evaluate this equation for each possible value of the trend.

The likelihood $p(obs|trend)$.

- This is simple (in principle), just the observation errors!

$$obs = H(Nature) + obs\ \ error$$

- Often observation errors are Gaussian, i.e.

$$obs\ \ error \sim N(0, \sigma^2)$$

- We need $p(obs|trend)$ for each value of $trend$
- Each of these values is given, so it acts as *Nature*, so we find

$$p(obs|trend) \propto \exp\left[-\frac{1}{2}\frac{(trend - obs)^2}{\sigma^2}\right]$$

The prior $p(trend)$.

- If you know something about the trend from other sources, e.g. a best value and its uncertainty, than that defines the prior
- If you don't know anything … that never happens
- Flat prior, Jeffrey's prior, symmetry arguments,…
- The point is that you have to say explicitly what your prior is.
- That way it becomes science because now we can debate about it !

So the machine should calculate

$$p(trend|obs) = \frac{p(obs|trend)}{p(obs)} p(trend)$$

- This is the full answer, a pdf of the trend given the observations
- No yes/no
- No hypothesis
- No difficult thinking

That’s all folks!

And what about significance testing???

What is significance testing?

We are taught to infer the truth of the hypothesis H_0 that the trend is zero :

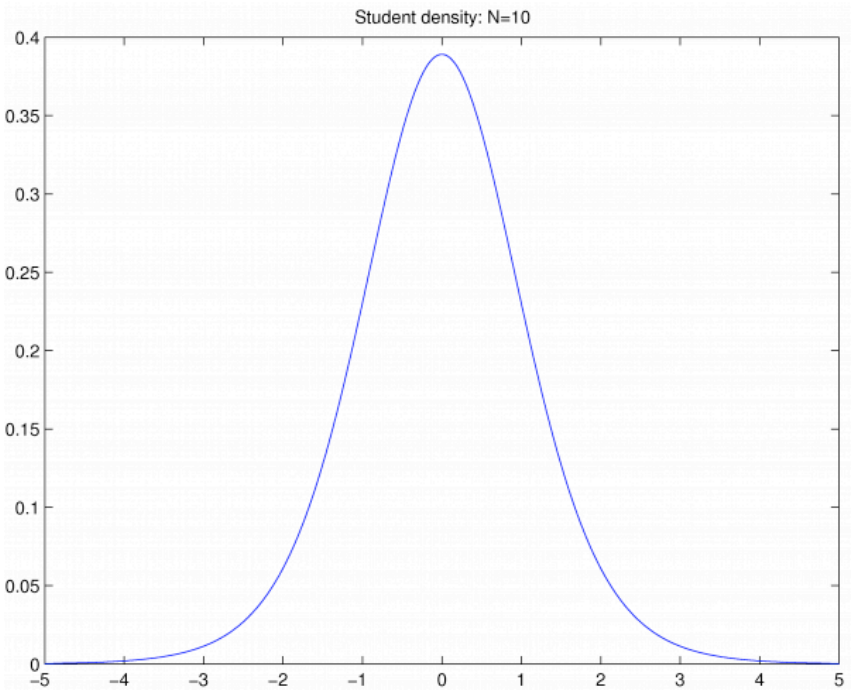
$$H_0 : \beta_0 = 0$$

Find a test statistic T with known distribution that can be computed from measurements.

We need a hypothetical distribution generated from an infinite number of observation sets …

$$T = \frac{\beta - \beta_0}{\sqrt{S^2}}$$

Lets assume T is Student's t



What is significance testing?

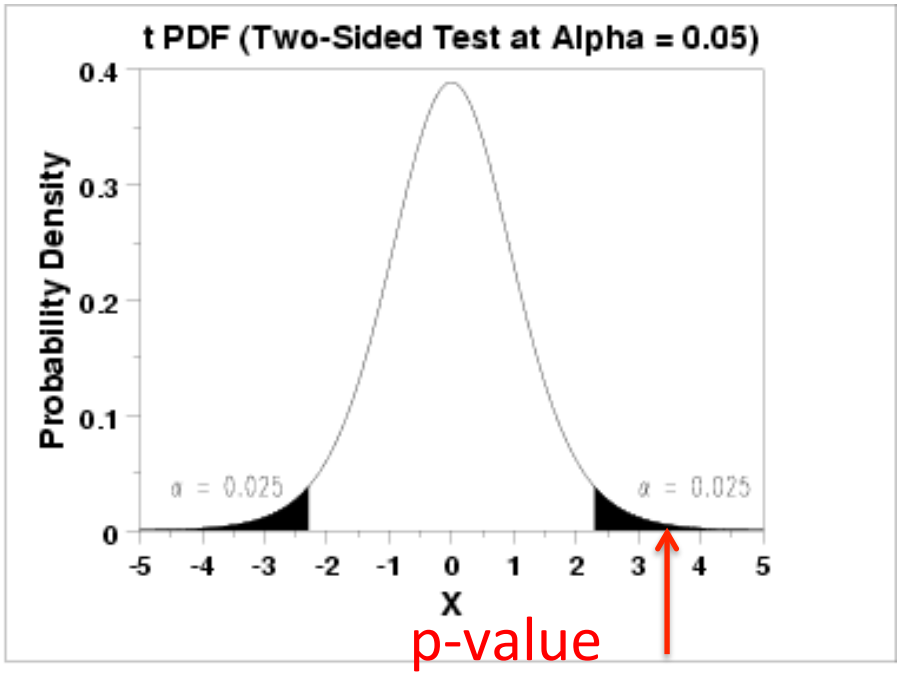
- Define a significance level α, as the probability threshold above which H_0 will be rejected: $\alpha = Pr(T \in R_{reject}|H_0)$
  Let's use … $\alpha = 0.05$

- Find the region for T for which H is rejected, so solve for R_{reject} in

$$Pr(T \in R_{reject}|H_0) = 0.05$$

- Determine if p=T_{obs} falls in that region, so if p is in R_{reject}

- If it does H_0 is rejected, if not, H_0 is not rejected.



What is significant testing?

Luckily we didn't

Fail to reject the null hypothesis

That would be a triple negative…

Who invented this jargon???

A few questions…

- How do we find the distribution of T?
- What should we choose for α? That seems to depend mainly on what others in the field have done, so α=0.05… (where is the science?)
- Is a yes/no answer useful?
- …

The full critics… 1

1) We want to know:

What is the probability of different trends given the observations.

Statisticians force us to ask questions in term so hypothesis, so how likely is the hypothesis given the observations, $Pr(H_0|T_{obs})$

Is that really what we want to know???

2) However, the test they invert works with $Pr(T \in R_{reject}|H_0)$, so the 'inverse'.

And these are REALLY different…

What is wrong with the following reasoning?

(Read Maarten Ambaum, J Climate, 2010)

1) My trend estimate stands out from the noise.
2) So my trend estimate is not likely to be caused by noise.
3) It is therefore unlikely that what I am seeing is noise.
4) The trend estimate is therefore positive evidence that there is really something happening.
5) This provides evidence for my theory.

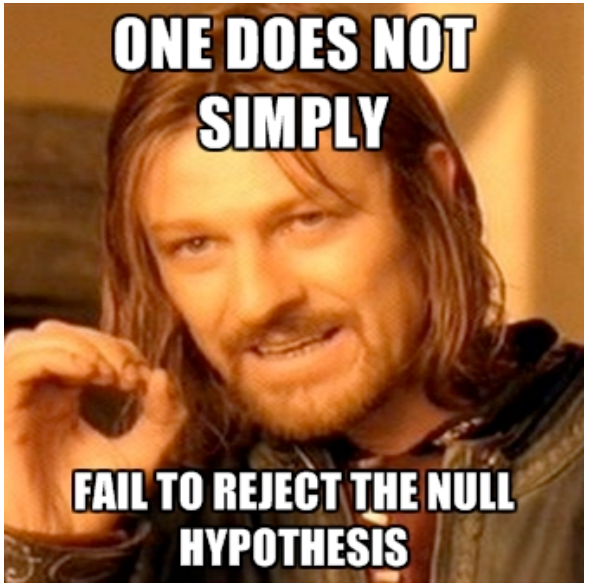
2) *Pr (I observe something | My system just produces noise)* is low

3) *Pr (My system just produces noise | I observe something)* is low

These are different. We want 3), but significant testing provides 2).

The full critics… 2

3) The hypothesis is always false, as the trend is never exactly zero! So unlike the original problem statement suggests, the truth of the hypothesis is not tested.



4) The true/false dichotomy is about decision making, not about science. What if the observations are close to the edge of the decision region?

The full critics 3

5) It doesn't tell is about the size of the effect, only yes or no.

6) The word 'significant' is misused, it should mean how big the effect is and if it can be replicated. Not so here…

7) It is widely misunderstood. It does not tell us if a hypothesis is true, it doesn't tell how likely the hypothesis is, etc etc.

8) It doesn't use any prior information, so no systematic scientific progress. (Also, results often need to be tempered after the fact by prior knowledge…)

The full critics 4

9) It is open to easy abuse by selection. Statistically, too many articles present significant results, leading to a bias. If not statistically significant it is harder to publish a result.

10) Only one hypothesis can be H_0, so the test is asymmetric.

11) It puts strong restrictions on sample size. A number of small sample-size studies might tell us more than one large study, but none of them will pass a significant test individually.

12) It emphasis random errors in large samples, instead of errors in individual measurements, so these errors tend to be ignored, while we should address them.

The full critics 5

13) It requires a number of unjustified assumptions: the distribution of the test statistic, sampling has to be purely random, α-value highly sensitive to scale transformations.

14) The results of two studies cannot be compared as random differences become highly amplified.

15) It is easy to play dirty, so choose the α-value after finding the p-value such that your result is significant…

The full critics 6

16) Fisher himself didn't see it as a good test, but simply as a quick and dirty first check to see if further research was necessary…

So this way of hypothesis testing just doesn't work for real science…

Conclusions

- As scientists we are interested the probability of the trend given a set of observations
- We know how to calculate that
- It is a learning framework, so new observations can easily be taken into account
- Statistical significance testing:

1. gives the wrong answer to the wrong question
2. is hopelessly complicated

So there is work to do:

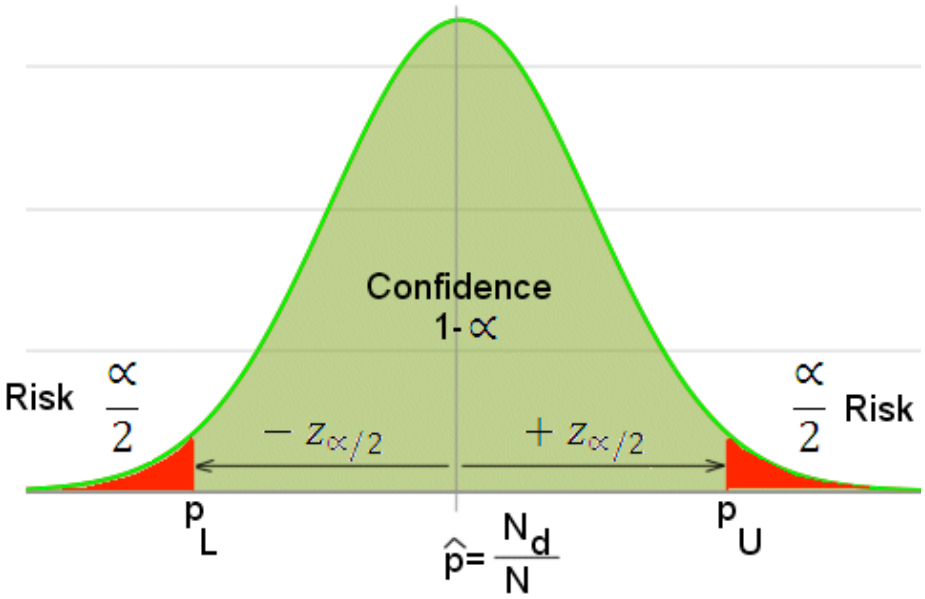
- Never use significance testing again !
- Reject papers that use it !
- Remove it from the curriculums !

And confidence intervals?

1) Determine the distribution of a test statistic T, e.g T is the (normalised) difference in the mean, so Student's t distributed.
2) Set α and define a confidence interval $Pr(-c(\alpha) < T < c(\alpha)) = 1-\alpha$,
3) Do the experiment and calculate the boundaries of the interval from the experimental mean and variance, around the mean.

This is often interpreted as:

There is a change of $(1-\alpha)$ that the true mean lies in this interval.



Confidence intervals???????

However, that is an incorrect interpretation. The correct frequentist interpretation is:

After numerous identical experiments 95% of the calculated confidence intervals will contain the true parameter value.

Each interval in isolation has a 0% or a 100% probability of containing the true parameter value…

Is this useful???