MSA101/MVE187 2021 Lecture 1 Introduction to Bayesian statistics

Petter Mostad

Chalmers University

August 30, 2021

Today's contents

- What to expect in the course
- The Bayesian approach to statistics: Introduction and overview
- The classical statistical approach: A comparison

What do I expect from you?

- Formal expectations:
  - Three individual obligatory assignments
  - A final written exam, determining the grade
- In addition, my actual expectations:
  - Read up on literature BEFORE lectures.
  - Be active in connection with lectures. Ask questions! (IRL / via zoom / afterwards)
  - Make sure you do exercises that help YOU learn. Take advantage of the exercise sessions.

What can you expect from the course?

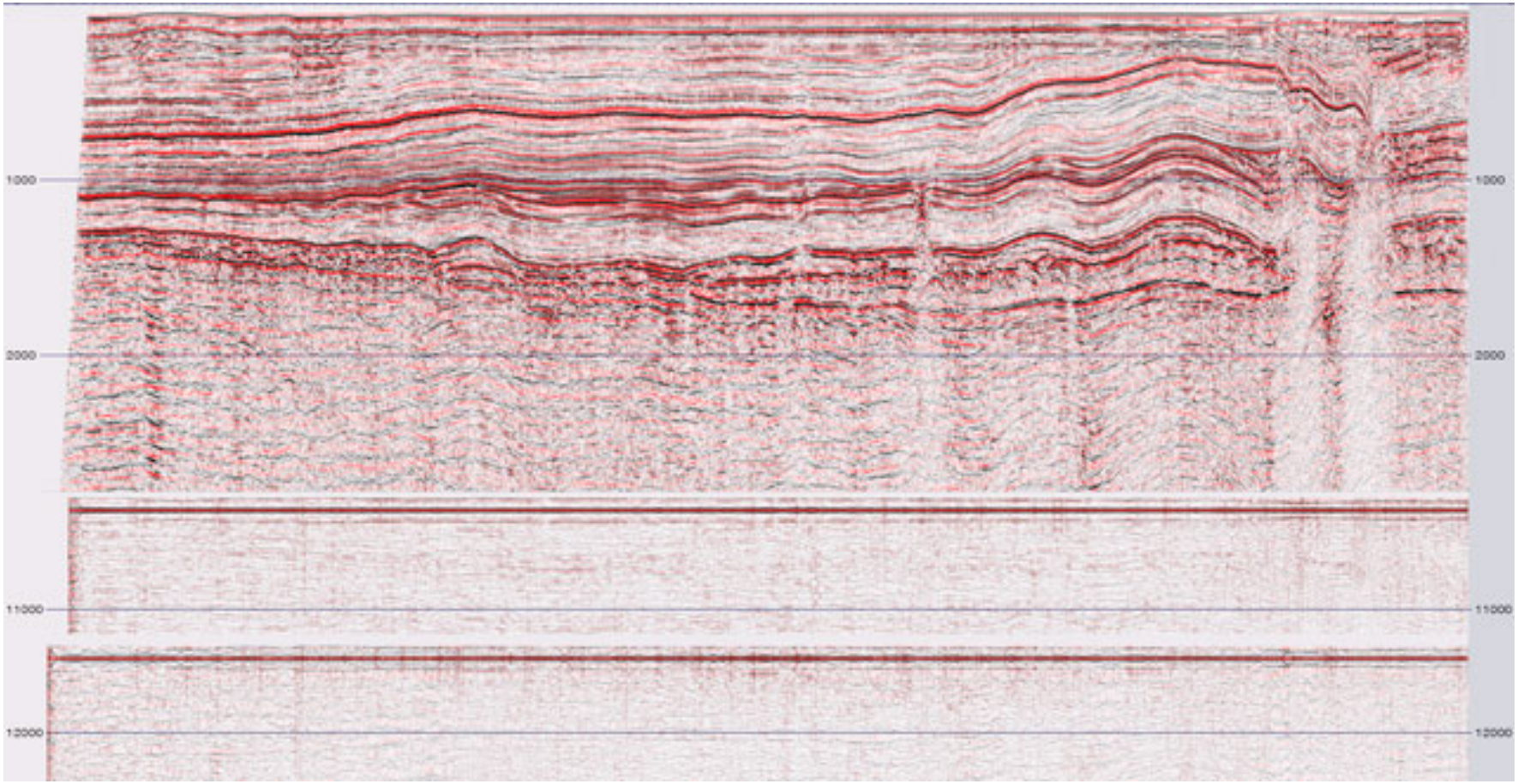
- A Canvas course page, also used for handing in assignments.
- Two lectures each week: In the fifth week the lecturer is Umberto Picchini.
- One exercise session each week: Helping YOU work.
- Outside lectures and exercise sessions I will answer mail (and Canvas messages) when I have time.

THE IDEA BEHIND THIS COURSE:
A statistics course where we by default use the Bayesian approach, and focus on methods to facilitate this approach.

How to do Bayesian statistics

- Decide on
  - what you want to learn about, formulated as as a *variable* Y_{pred} with one or more dimensions: *What do you need to know to support your decisions?*
  - What data will you use, formulated as a variable Y_{data} with usually many dimensions.
- Decide on a probability model, assigning a probability (or probability density) to all possible combinations of values for Y_{pred} and Y_{data}.
- Base your decisions on the conditional probability distribution on Y_{pred} obtained by fixing Y_{data} to the observed value.

Example: Deciding where to drill an oil well



Simplest example: Flippling a biased coin

- Y_{pred} will be H or T. Question: What is probability of H in next throw?
- Unbiased coin: trivial
- Biased coin: Need data! Y_{data}: results of previous throws, e.g., $Y_{data} = HTTHTTT$.
- (First) model: Probability of heads is either 0.7 or 0.3, with a probability 0.5 for each possibility.
- Probability of observing a sequence of r heads in N throws:

$$0.5 \cdot 0.7^r \cdot (1 - 0.7)^{N-r} + 0.5 \cdot 0.3^r \cdot (1 - 0.3)^{N-r}$$

- If, say $Y_{data} = HTTHTTT$ and $Y_{pred} = H$ we can compute

$$\Pr(Y_{pred} \mid Y_{data}) = \frac{\Pr(Y_{data}, Y_{pred})}{\Pr(Y_{data})} = \frac{\Pr(HTTHTTTH)}{\Pr(HTTHTTT)}$$

$$= \frac{0.5 \cdot 0.7^3 \cdot 0.3^5 + 0.5 \cdot 0.3^3 \cdot 0.7^5}{0.5 \cdot 0.7^2 \cdot 0.3^5 + 0.5 \cdot 0.3^2 \cdot 0.7^5} = 0.3291892$$

- Exact same result if Y_{data} is instead number of heads in N tries, ignoring sequence.

Biased coin example

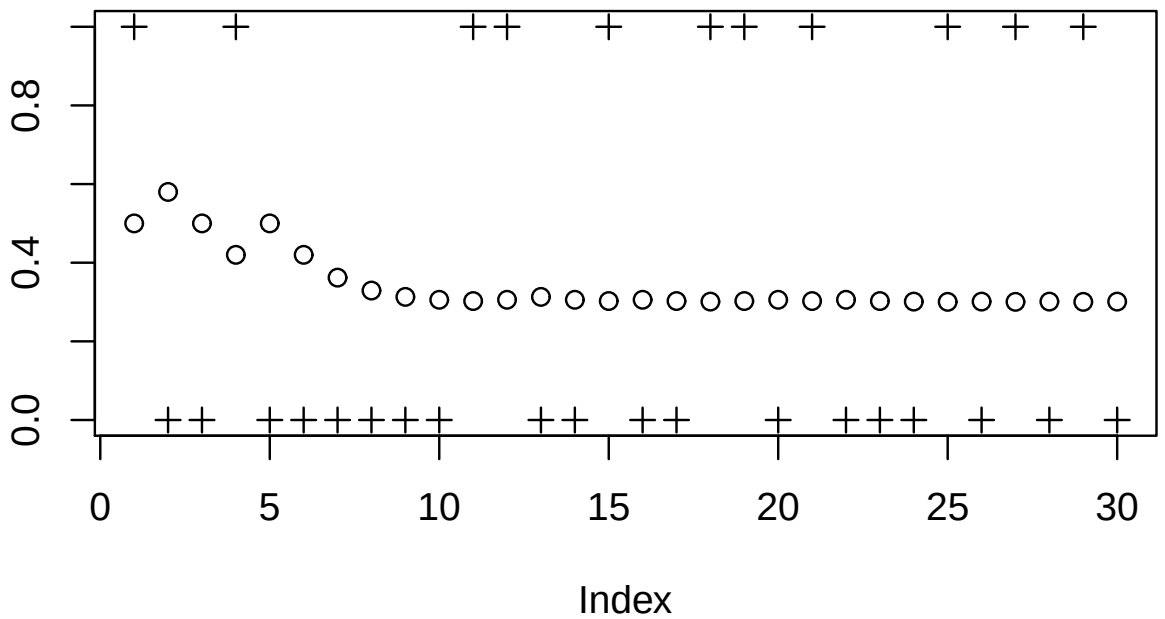


Figure: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. Model: The probability θ of heads is either 0.7 or 0.5, with $\Pr(\theta = 0.7) = \Pr(\theta = 0.3) = 0.5$.

A note on notation

- In Bayesian statistics, **probability density functions** and **probability mass functions** are more useful than cumulative distributions.
- When X is a discrete random variable, we write $\pi(X)$ or $\pi(x)$ for its probability mass function (assuming distribution clear from context).
- When X is a continuous random variable with a probability density function, we **also** write $\pi(X)$ or $\pi(x)$ for its probability density function (assuming the distribution is clear from the context).
- This extends to conditional distributions: $\pi(x \mid y)$ is a conditional probability mass function if x is discrete, and a conditional density if x is continuous.
- Thus we can write general formulas, like Bayes formula, like this:

$$\pi(x \mid y) = \frac{\pi(y \mid x)\pi(x)}{\pi(y)}$$

- We use integrals to mean either integration or summation, depending on whether the variable is discrete or continuous, e.g., $\int \pi(x, y)\, dy = \pi(x)$.

Formulation of models using a parameter θ

- Use a parameter (vector) θ so that
  - You specify a joint distribution on $(Y_{data}, Y_{pred}, \theta)$ so that its marginal (summing or integrating out θ) is the intended distribution on (Y_{data}, Y_{pred}).
  - Y_{data} and Y_{pred} are *conditionally independent* given θ. In other words,

$$\pi(Y_{pred} \mid \theta, Y_{data}) = \pi(Y_{pred} \mid \theta)$$

- The joint distribution is generally specified by specifying
  - $\pi(\theta)$ **called the prior**
  - $\pi(Y_{data} \mid \theta)$ **called the likelihood**
  - $\pi(Y_{pred} \mid \theta)$
- We then have $\pi(Y_{data}, Y_{pred}, \theta) = \pi(\theta)\pi(Y_{data} \mid \theta)\pi(Y_{pred} \mid \theta)$.
- $\pi(\theta \mid Y_{data})$ is **called the posterior**
- $\pi(Y_{pred} \mid Y_{data}) = \int \pi(Y_{pred}, \theta \mid Y_{data})\, d\theta =$
  $\int \pi(Y_{pred} \mid \theta, Y_{data})\pi(\theta \mid Y_{data})\, d\theta = \int \pi(Y_{pred} \mid \theta)\pi(\theta \mid Y_{data})\, d\theta$

Returning to our biased coins

- In our biased coin example, the model can be reformulated with θ representing the probability of heads.
- In our first model, θ is a discrete variable with the possible values 0.7 and 0.3, and prior $\pi(\theta = 0.7) = \pi(\theta = 0.3) = 0.5$.
- If Y_{data} and Y_{pred} specify counts of heads, we have $Y_{data} \mid \theta \sim \text{Binomial}(N, \theta)$ and $Y_{pred} \mid \theta \sim \text{Binomial}(1, \theta)$.
- Compute the posterior for θ, and the predictive distribution for Y_{pred} given Y_{data}!
- An alternative model uses that θ is any real value in $(0, 1)$, with a uniform prior. Then $\pi(\theta) = 1$.
- We show next time that the posterior for θ now has a Beta distribution, while the distribution of Y_{pred} given Y_{data} gets a Beta-Binomial distribution.

Biased coin example

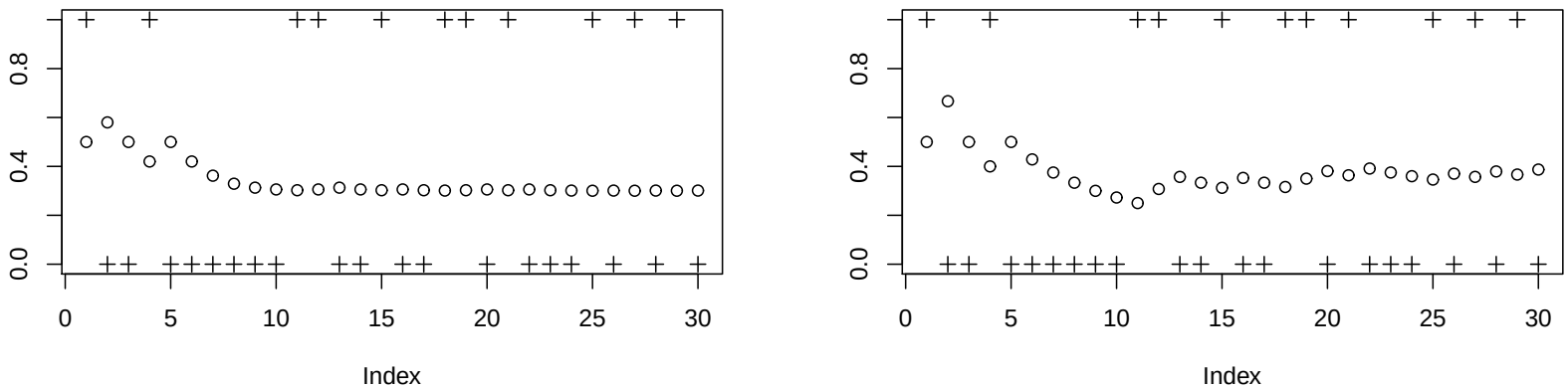


Figure: The probability of heads at each point in a sequence of observations, or the probability of "success", conditioning on the previous observations. The priors used are $\pi(\theta = 0.7) = \pi(\theta = 0.3) = 0.5$ (left) and $\theta \sim \text{Uniform}(0, 1)$ (right).

How you (may) do classical statistics

- Decide on
  - what you want to learn about, formulated as as a *variable* Y_{pred} with one or more dimensions: *What do you need to know to support your decisions?*
  - What data will you use, formulated as a variable Y_{data} with usually many dimensions.
- Decide on a parameter θ, and specify
  - A likelihood $\pi(Y_{data} \mid \theta)$
  - $\pi(Y_{pred} \mid \theta)$

  so that you have conditional independence:
  $\pi(Y_{pred} \mid \theta, Y_{data}) = \pi(Y_{pred} \mid \theta)$.
- **You do not specify a prior** $\pi(\theta)$
- Instead, you specify a **method for estimation** to derive $\hat{\theta}$ from the data Y_{data} and the likelihood $\pi(Y_{data} \mid \theta)$.
- Base your decisions on the probability distribution $\pi(Y_{pred} \mid \hat{\theta})$.

First difference: Using a fixed $\hat{\theta}$ instead of a posterior distribution $\pi(\theta \mid Y_{data})$

- In our biased coin example: If you *know* the coin has either 0.7 or 0.3 probability for heads, following the procedure above means
  - Using Y_{data} to *decide* on either 0.7 or 0.3 as a value for $\hat{\theta}$.
  - Using that estimate for predicting Y_{pred}
- In a linear regression setting:
  - You estimate a line from the data, for example using least squares.
  - You use this line for predictions.
- In general: Instead of using

$$\int \pi(Y_{pred} \mid \theta)\pi(\theta \mid Y_{data})\, d\theta$$

(the Bayesian approach) you use

$$\pi(Y_{pred} \mid \hat{\theta})$$

NOTE: If $\pi(\theta \mid Y_{data})$ is closely concentrated around $\hat{\theta}$ these two approaches give (almost) the same thing!

Problems with "throwing away uncertainty"

- If you "forget" about your uncertainty around your knowledge of θ, you may become overconfident in your predictions for Y_{pred}.
- As a response, classical methods may derive $\hat{\theta}$ **together with** an uncertainty for this estimate.
- If the uncertainty is represented in a density $f(\theta)$, the prediction may be based on

$$\int \pi(Y_{pred} \mid \theta) f(\theta)\, d\theta$$

- Example: This is used for predictions in basic linear regression.

Updated: How you might do classical statistics

- Decide on
  - what you want to learn about, formulated as as a *variable* Y_{pred} with one or more dimensions: *What do you need to know to support your decisions?*
  - What data will you use, formulated as a variable Y_{data} with usually many dimensions.
- Decide on a parameter θ, and specify
  - A likelihood $\pi(Y_{data} \mid \theta)$
  - $\pi(Y_{pred} \mid \theta)$

  so that you have conditional independence:
  $\pi(Y_{pred} \mid \theta, Y_{data}) = \pi(Y_{pred} \mid \theta)$.
- **You do not specify a prior** $\pi(\theta)$
- Instead, you specify a **method for estimation** to derive $\hat{\theta}$ from the data Y_{data} and the likelihood $\pi(Y_{data} \mid \theta)$, **and** you derive an uncertainty density for your estimate $f(\theta)$.
- Base your decisions on the probability distribution
  $\int \pi(Y_{pred} \mid \theta) f(\theta)\, d\theta$.

Still problems, e.g, when relying on data from several sources

- In more advanced, realistic applications, we learn about the parameters θ using data from several sources. (Example: Oil reservoir!)
- How do we combine the estimates, and their uncertainties? Classical statistics does not provide a clear answer.
- In Bayesian statistics, you can string the analyses together, with the posterior from one becoming the prior for the next. (More on this next time).

Second difference: Specifying an estimation method instead of a prior

- Classical statistics seems to have a huge advantage in not needing a prior!
- "Priors are subjective, and thus not scientific, whereas estimation methods can be derived objectively"
- ...is this true?
- Problem: An estimation method depends both on an estimator and a model formulation, and whether an estimator is objectively good may vary with equivalent model formulations.

Unbiased estimators

- Given a likelihood $\pi(Y_{data} \mid \theta)$, an *unbiased estimator* is a function g from possible values for the data to possible values for θ, such that, for all θ,

$$\mathrm{E}_{Y_{data}|\theta}\, g(Y_{data}) = \theta$$

- Unbiasedness is regarded as an "objective" way to say that an estimator is "good".

Example

- Assume we have a sequence of independent trials each resulting in success (1) or failure (0), with a probability of succes equal to p. Assume we have observed the following data:

$$0, 1, 0, 0, 1, 0, 0, 1$$

  We then make the estimate $3/8 = 0.375$ for p. How "good" is this estimate? Is it unbiased?
- It depends on which model formulation and which estimator we have used!
- Alternative 1: The estimator is: Make 8 trials, let X be the number of successes, and compute $\hat{p} = X/8$.
- Alternative 2: The estimator is: Make trials until you have produced 3 successful trials, let X be the number of trials you needed to do, and compute $\hat{p} = 3/X$.

Continuation of example

- Exercise: Prove that the estimator in alternative 1 is unbiased (easy), and that the estimator in alternative 2 is biased (more difficult).
- Our point here: If we use the biasedness of the *estimator* to judge whether the *estimate* 0.375 is good, the result depends on which estimator we are using, which depends on what went on in the head (the plans) of the person doing the experiments.

Continuation of example

- In the same situation as above, and the same observations, we want to make a hypothesis test with $H_0 : p \geq 0.6$, and alternative hypothesis $H_1 : p < 0.6$. What is the p-value?
- To answer the question, we need to know which *test statistic* should be used.
- Alternative 1: The test statistic is: Make 8 trials and let X be the number of successes. Then, assuming $p = 0.6$, we get $X \sim \text{Binomial}(8, 0.6)$. The possible values for X and their probabilities are

0	1	2	3	4	5	6	7	8
0.001	0.008	0.041	0.124	0.232	0.279	0.209	0.090	0.017

We get that the p-value becomes 0.174; the sum of the probabilities for $X = 0, 1, 2, 3$.

Continuation of example

- Alternative 2: The test statistic is: Make trials until 3 successes have appeared and let X the number of trials necessary. Then, assuming $p = 0.6$, we get $X \sim$ Neg-Binomial$(3, 0.6)$. The possible values for X and their probabilities are

3	4	5	6	7	8	9	10	11
0.216	0.259	0.207	0.138	0.083	0.046	0.025	0.013	0.006

12	13	14	15	16,17,. . .
0.003	0.001	0.001	0.000	total 0.000

  We get the p-value 0.095; the sum of the probabilities for $8, 9, 10, \ldots$.

- Note that if we use a significance level of 0.1, we will reject the null hypothesis using the second test statistic, but not using the first test statistic.

Correspondence between Bayesian and classical methods

- So maybe there are "reasonable" and "unreasonable" estimators, as well as "reasonable" and "unreasonable" priors?
- If a classical approach derives an estimate $\hat{\theta}$ as well as a distribution $f(\theta)$ for the uncertainty of the estimate, can we find a prior $\pi(\theta)$ so that the posterior $\pi(\theta \mid Y_{data})$ will be the same as $f(\theta)$, for all choices of the data?
- In some cases we can.

Example: The normal model with fixed variance

- Assume $Y_{data} = (y_1, y_2, \ldots, y_n)$ where the y_i are real values, and that, independently given θ,

$$y_1, y_2, \ldots, y_n, y_{pred} \sim \text{Normal}(\theta, 1)$$

  Write $\overline{y} = (y_1 + \cdots + y_n)/n$.
- We show (next time): If the prior is $\pi(\theta)$, the posterior for θ is proportional to[1] $\text{Normal}(\theta, \overline{y}, 1/n)\pi(\theta)$.
- If we use a "flat prior" for θ (discussed next time), the posterior density becomes $\text{Normal}(\theta; \overline{y}, 1/n)$.
- In classical statistics we use the unbiased estimator $\hat{\theta} = g(Y_{data}) = (y_1 + y_2 + \cdots + y_n)/n$. Its distribution is

$$g(Y_{data}) \sim \text{Normal}(\theta, 1/n).$$

- Although NOT FORMALLY CORRECT, we might use the density $f(\theta) = \text{Normal}(\overline{y}, 1/n)$ to indicate the uncertainty in the estimate. Then same result as the Bayesian analysis!

[1] $\text{Normal}(\theta; \overline{y}, 1/n)$ means the value at θ of the density of the Normal distribution with expectation $\overline{y}$ and variance $1/n$

Example, with numerical values

Example:

- In the example above, the observed values are 4.2, 5.6 and 4.6.
- The estimate for θ becomes 4.8, the mean of the numbers.
- A 95% confidence interval for θ can then be computed as

$$\left[4.8 - 1.96 \cdot \frac{1}{\sqrt{3}}, 4.8 + 1.96 \cdot \frac{1}{\sqrt{3}}\right] = [3.67, 5.93]$$

- The correct interpetation of the interval: If three numbers are resampled from the distribution many times, the re-computed confidence intervals will contain θ with probability 95%.
- Another common interpretation: The interval [3.67, 5.93] contains θ with 95% probability. INCORRECT in classical statistics as θ is then not a random variable.
- However, in Bayesian statistics, we have $\theta \mid Y_{data} \sim \text{Normal}(\overline{y}, 1/n)$ and the statement above becomes CORRECT.

Philosophical differences: What does probability mean when applied in the real world?

- The mathematical theory of probability is not under discussion.
- What does it *mean* when we say:
  - The probability of yahtzee (5 equal dice) in one throw is 0.00077.
  - The probability of rain tomorrow is 0.3
  - The probability that this oil well will produce oil is 0.93.
- Classical focus: *Repeatable events*
- Bayesian approach: Making probability models for *knowledge* about some part of the real world, not the part of the real world itself.

Summary (a personal view)

WARNING: Contains Bayesian propaganda!

- The Bayesian paradigm can in principle be applied in all contexts where probabilistic predictions are wanted.
- Advantage: Work clearly divided into model specification ("subjective") and making predictions (mathematical computations).
- The classical approach is NOT more "objective" than the Bayesian approach.
- All statistical modelling require decisions based on contextual information, outside of the data. In specifying the prior, the Bayesian method makes very clear what this information is and how it is used. The classical method less so.
- I believe: All statistical methodology that behaves sensibly can be viewed as (an approximation to) a Bayesian approach, and a particular choice of prior.
- Ideas in classical statistics can generally be translated to ideas in Bayesian statistics (example: bias-variance tradeoff).
- Remember: One should model *knowledge* about the real world, not the real world itself.

The academic discussion; history

- Bayesian statistics is named after rev. Thomas Bayes who formulated a version of Bayes' theorem in 1763.
- Early probabilists, such as Laplace, worked in ways compatible with the Bayesian paradigm.
- In the 20'th century, the frequentist paradigm dominated, developed for example by Fisher.
- Towards the end of the 20'th century, there was a furious academic discussion, between "Frequentists" and "Bayesians".
- Fast computers facilitated the rise of Bayesian statistics in practice.
- Today, a lot of basic courses still focus on Frequentist methods, whereas applied research can often be Bayesian or "agnostic" (i.e., "anything goes").