MSA101/MVE187 2022 Lecture 1
Introduction to Bayesian statistics

Petter Mostad

Chalmers University

August 29, 2022

Following textbooks in classical statistics: Unbiased estimators

Assume x (maybe $x = (x_1, \ldots, x_n)$) has some probability distribution with parameter θ.

- Can be specified by some probability density function $f(x; \theta)$.
- Fixing x, $f(x; \theta)$ as a function of θ is the *likelihood function*.
- An *estimator* for θ is a function $g(x)$ that "estimates" θ; an *estimate* is the value of g when x is equal to observed data.
- An *unbiased estimator* has the property

$$\mathrm{E}_{x|\theta}\left[g(x)\right] = \theta.$$

- An estimator is considered "good" if it is unbiased.

Problems with classical framework: Example 2

- 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 $\hat{p} = 3/X$.

Problems with classical framework: Example 2

- 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.

Problems with classical framework: Example 3

- 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 $X = 0, 1, 2, 3$.

Problems with classical framework: Example 3

- 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.

Basic ideas of Bayesian analysis

- Represent all observed data, unobserved data, *and parameters* as random variables, and consider a joint probability distribution for all these random variables.
- *We need to accept using "probability densities" that do not integrate to 1! Such densities are called* **improper densities**.
- *Learning* corresponds to considering *conditional densities*!
- No mention of estimation; no need for this concept in Bayesian statistics!
- Focus on *prediction*: What you want to predict should be represented with a random variable Y_{pred} in the joint probability distribution. Find $Y_{\text{pred}} \mid \text{data}$.

Biased coin example

- When flipping coin with known probability of heads (H): prediction is trivial.
- Unknown probability of heads (possible bias): Predict outcome of next throw Y_{pred} to be H or T given throws so far: $Y_{\text{data}} = HTTHTTT$.
- 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:

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

- We can compute for example

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

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

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

Biased coin example

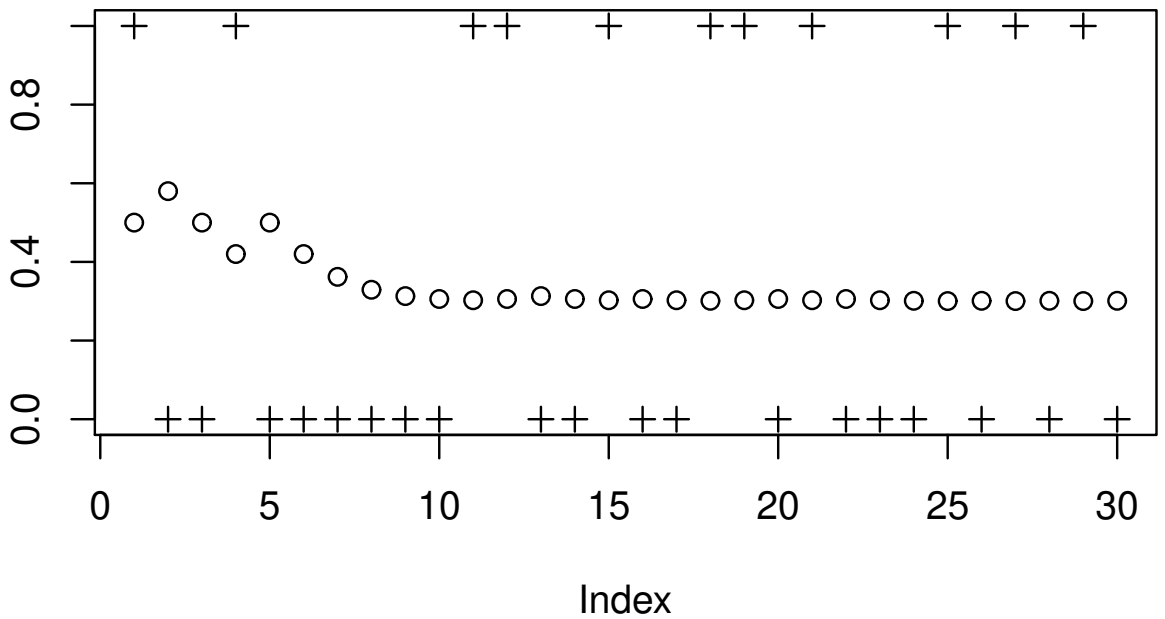


Figure: The probability of heads at each point in a sequence of observations, conditioning on the previous observations of heads and tails. Model: The probability θ of heads is either 0.7 or 0.5, with
$\Pr(\theta = 0.7) = \Pr(\theta = 0.3) = 0.5$.

Biased coin example

- Reformulating with θ representing the probability of heads:

$$\begin{aligned} \theta & \quad \text{discrete variable with } \Pr(\theta = 0.7) = \Pr(\theta = 0.3) = 0.5 \\ Y_{\text{data}} \mid \theta & \sim \text{Binomial}(N, \theta) \\ Y_{\text{pred}} \mid \theta & \sim \text{Binomial}(1, \theta) \end{aligned}$$

- We get *in general with conditional independence of Y_{data} and Y_{pred} given θ*:

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

- In our case:

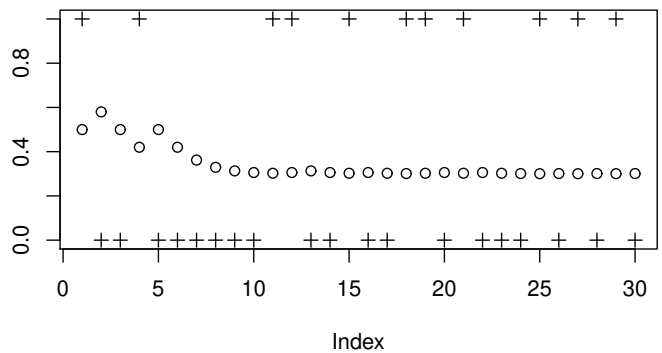
$$\begin{aligned} \Pr(\theta = 0.7 \mid Y_{\text{data}}) &= 0.073 \\ \Pr(\theta = 0.3 \mid Y_{\text{data}}) &= 0.927 \end{aligned}$$

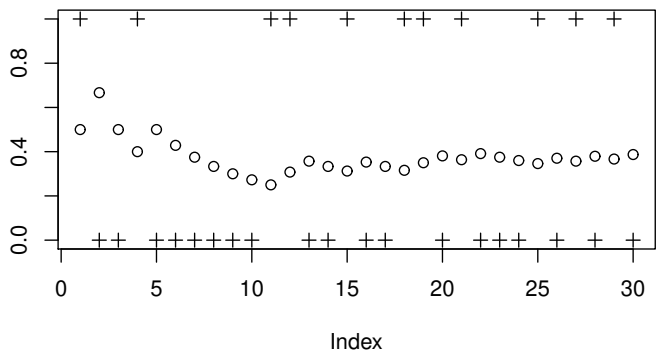
- In our case:

$$\begin{aligned} & \Pr(Y_{\text{pred}} = H \mid Y_{\text{data}}) \\ = \; & \Pr(Y_{\text{pred}} = H \mid \theta = 0.7) \cdot 0.073 + \Pr(Y_{\text{pred}} = H \mid \theta = 0.3) \cdot 0.927 \\ = \; & 0.3292 \end{aligned}$$

Biased coin example

- 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 is a Beta distribution, while the distribution of Y_{pred} given Y_{data} is a Beta-Binomial distribution.





- 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).

Bayesian statistics (summary)

- Formulate a joint probability density model

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

- (In classical stats: No prior $\pi(\theta)$ is formulated.)
- Find the posterior distribution $\pi(\theta \mid Y_{\text{data}})$.
- (In classical stats: Find estimate $\hat{\theta}$.)
- Make probabilistic predictions using

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

  (or $\int \pi(Y_{\text{pred}} \mid \theta, Y_{\text{data}})\pi(\theta \mid Y_{\text{data}})\, d\theta$ if necessary).
- (In classical stats, predictions are often made using $\pi(Y_{\text{pred}} \mid \hat{\theta})$.)

Discussion: Bayes vs. frequentist!

- **Frequentists**: Priors are subjective, WE however are SCIENTISTS and use only data!
- **Bayesians:** Your methods do not answer the central applied questions. Bridging this gap IMPLICITLY adds the same information as that in a choice of prior.
- **Bayesians:** YOUR methods throw away uncertainty!
- **Frequentists**: We are perfectly capable of propagating uncertainty, WHEN NEEDED.
- **Bayesians:** Yeah, but where does your uncertainty come from? Confusing the uncertainty of an estimator with the uncertainty in a parameter creates problems.
- **Frequentists**: Well, Bayesian statistics is not doable anyway: How do you find your prior in practice, and how do you compute your posterior and your predictions?
- **Bayesians**: Yes, finding a prior IS difficult in practice. But concerning the computational stuff, we have made huge progress since the 90's.

More arguments for a Bayesian point of view

- Bayesian work is clearly divided into model specification (using contextual information) and making predictions (mathematical computations). Applied classical statistics must also use contextual information, but how it is used can be less clear.
- Bayesian modelling provides a choice between using priors that reflect little knowledge ("non-informative priors", whatever that means) and priors that contain important information one wants to use, even if it is not called "data".
- There is often a way to translate between results and concepts in classical and in Bayesian statistics.
- May issues in classical statistics, for example overfitting or multiple testing issues, become much easier to handle when translated to a Bayesian context (in the opinion of Bayesians).

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.

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").

Final example of connection between Bayesian and classical statistics

Assume random variables $x_1, \ldots, x_n$ have a probability distribution with a parameter θ.

- If we construct functions l_1 and l_2 so that

$$\Pr[l_1(x_1, \ldots, x_n) \leq \theta \leq l_2(x_1, \ldots, x_n)] = 0.95$$

  for the *random variables* $x_1, \ldots, x_n$ we say that $[l_1(x_1, \ldots, x_n), l_2(x_1, \ldots, x_n)]$ is a 95% **confidence interval** for θ.
- Interpretation: If we resample data from a model with parameter θ, newly computed intervals will cover θ with probability 95%.
- If we also have a probability distribution defined on θ, then an interval $[l_1, l_2]$ where l_1 and l_2 are *numbers* is called a 95% **credibility interval** for θ if

$$\Pr[l_1 \leq \theta \leq l_2 \mid x_1, \ldots, x_n] = 0.95.$$

- *However, when computed, the intervals can often be the same (if the prior on θ is chosen appropriately)*

Example: The normal model with unit variance

Assume $x_1, \ldots, x_n \mid \mu \sim \text{Normal}(\mu, 1)$

- Classical statistics: $\overline{x} \sim \text{Normal}(\mu, 1/n)$ and thus

$$\Pr\left[\overline{x} + 1.96\frac{1}{n} \leq \mu \leq \overline{x} + 1.96\frac{1}{n}\right] = 0.95$$

  where $\overline{x}$ is the random variable.
- We saw: If μ has a *flat prior* then $\mu \mid x_1, \ldots, x_n \sim \text{Normal}(\overline{x}, 1/n)$ and thus

$$\Pr\left[\overline{x} + 1.96\frac{1}{n} \leq \mu \leq \overline{x} + 1.96\frac{1}{n} \mid x_1, \ldots, x_n\right] = 0.95$$

  where μ is the random variable.
- NOTE: If the prior is $\pi(\theta)$ we get that $\pi(\mu \mid x_1, \ldots, x_n) \propto_\mu \text{Normal}(\mu; \overline{x}, 1)\pi(\mu)$ so the above does not hold unless $\pi(\mu) \propto_\mu = 1$.
- With observations 4.2, 5.6, and 4.6, we get $\overline{x} = 4.8$ and

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

  for both the confidence interval and the credibility interval!

What do I expect from you in this course?

- Formal expectations:
  - Three individual obligatory assignments
  - A final written exam, determining the grade
- In addition, my actual expectations:
  - Get familiar with the information on the Canvas course page.
  - Read up on literature BEFORE lectures.
    Student: "I think maybe students should be encouraged to skim through the relevant book chapters before the lectures because it really helped me when doing so."
  - Be active in connection with lectures. Ask questions!
  - Take responsibility for learning assumed background knowledge, such as running R and basic probability. But also ask me for help!
  - 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. Three lectures will be with 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.