Introduction to Discrete Bayesian Methods

Introduction to Bayesian Modeling

- In the social science researchers point of view, the requirements of traditional frequentistic statistical analysis are very challenging.
- For example, the assumption of normality of both the phenomena under investigation and the data is prerequisite for traditional parametric frequentistic calculations.

Continuous

age, income, temperature, ..

Discrete

FSIQ in the WAIS-III, Likert –scale, favourite colors, gender, ..

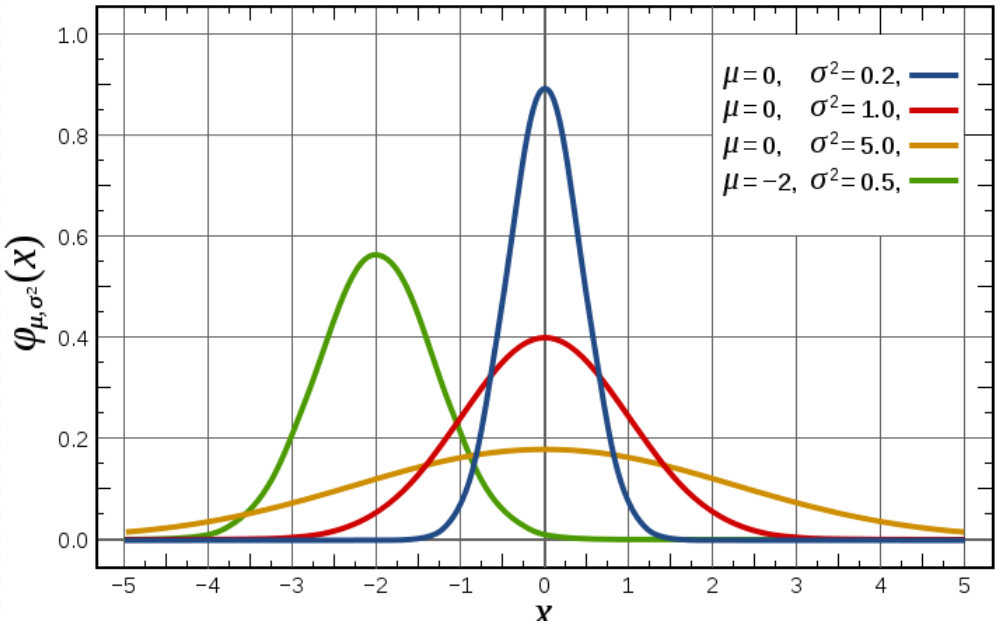
Introduction to Bayesian Modeling

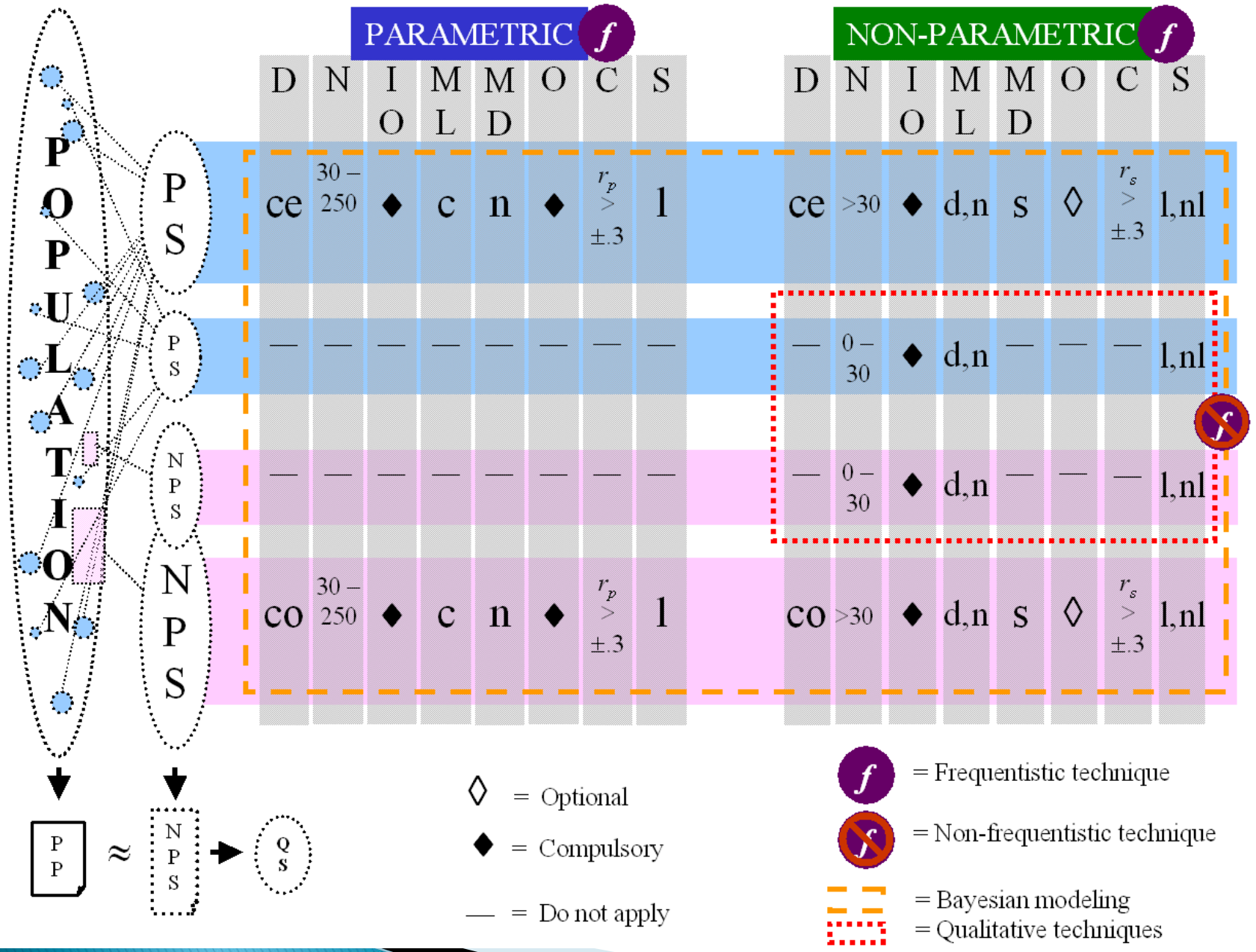
- In situations where
  - a latent construct cannot be appropriately represented as a continuous variable,
  - ordinal or discrete indicators do not reflect underlying continuous variables,
  - the latent variables cannot be assumed to be normally distributed,

  traditional Gaussian modeling is clearly not appropriate.
- In addition, normal distribution analysis sets minimum requirements for the number of observations, and the measurement level of variables should be continuous.

Introduction to Bayesian Modeling

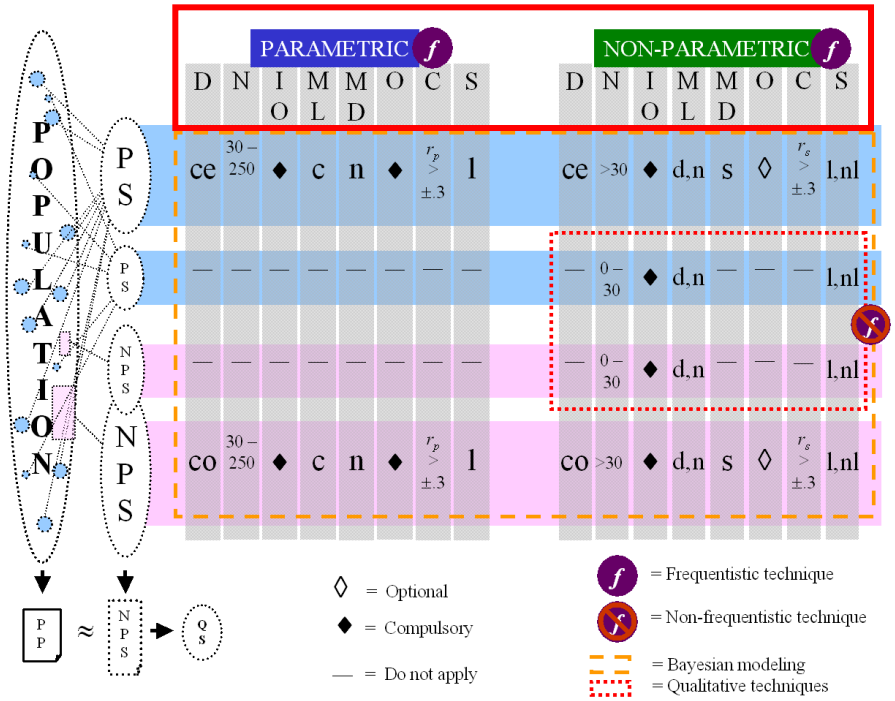
- Frequentistic parametric statistical techniques are designed for normally distributed (both theoretically and empirically) indicators that have linear dependencies.
  - Univariate normality
  - Multivariate normal
  - *Bivariate linearity*





(Nokelainen, 2008, p. 119)

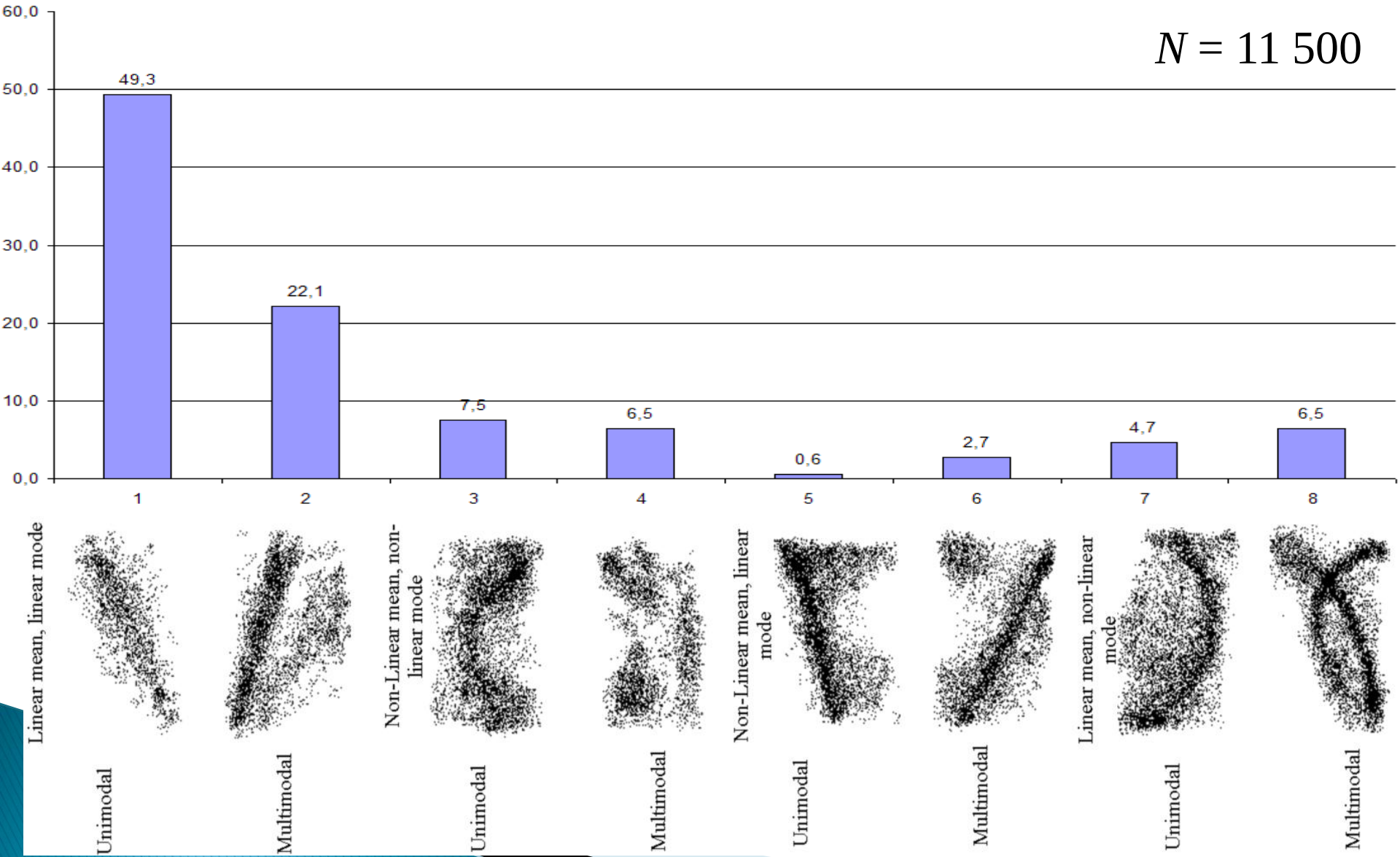
- The upper part of the figure contains two sections, namely “parametric” and “non-parametric” divided into eight sub-sections (“DNIMMOCS OLD”).
- Parametric approach is viable only if
  - 1) Both the phenomenon modeled and the sample follow normal distribution.
  - 2) Sample size is large enough (at least 30 observations).
  - 3) Continuous indicators are used.
  - 4) Dependencies between the observed variables are linear.
- Otherwise non-parametric techniques should be applied.



D = Design (ce = controlled experiment, co = correlational study)
N = Sample size
IO = Independent observations
ML = Measurement level (c = continuous, d = discrete, n = nominal)
MD = Multivariate distribution (n = normal, similar)
O = Outliers
C = Correlations
S = Statistical dependencies (l = linear, nl = non-linear)

Introduction to Bayesian Modeling

$N = 11\,500$



Introduction to Bayesian Modeling

Bayesian method

(1) is parameter-free and the user input is not required, instead, prior distributions of the model offer a theoretically justifiable method for affecting the model construction;

(2) works with probabilities and can hence be expected to produce robust results with discrete data containing nominal and ordinal attributes;

(3) has no limit for minimum sample size;

(4) is able to analyze both linear and non-linear dependencies;

(5) assumes no multivariate normal model;

(6) allows prediction.

Introduction to Bayesian Modeling

- Probability is a mathematical construct that behaves in accordance with certain rules and can be used to represent uncertainty.
  - The classical statistical inference is based on a frequency interpretation of probability, and the Bayesian inference is based on ”subjective” or ”degree of belief” interpretation.
- Bayesian inference uses conditional probabilities to represent uncertainty.
- $P(H \mid E,I)$ – the probability of unknown things or ”hypothesis” (H), given the evidence (E) and background information (I).

Introduction to Bayesian Modeling

- The essence of Bayesian inference is in the rule, known as **Bayes' theorem**, that tells us how to update our initial probabilities $P(H)$ if we see evidence E, in order to find out $P(H|E)$.

- A priori probability
- Conditional probability
- Posteriori probability

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E|H) \cdot P(H) + P(E|\sim H) \cdot P(\sim H)}$$

Introduction to Bayesian Modeling

- The theorem was invented by an english reverend Thomas Bayes (1701-1761) and published posthumously (1763).

Introduction to Bayesian Modeling

- Bayesian inference comprises the following three principal steps:

  (1) Obtain the initial probabilities $P(H)$ for the unknown things. (Prior distribution.)

  (2) Calculate the probabilities of the evidence E (data) given different values for the unknown things, i.e., $P(E \mid H)$. (Likelihood or conditional distribution.)

  (3) Calculate the probability distribution of interest $P(H \mid E)$ using Bayes' theorem. (Posterior distribution.)

- Bayes' theorem can be used sequentially.

Introduction to Bayesian Modeling

- If we first receive some evidence E (data), and calculate the posterior $P(H \mid E)$, and at some later point in time receive more data E', the calculated posterior can be used in the role of prior to calculate a new posterior $P(H \mid E,E')$ and so on.
- The posterior $P(H \mid E)$ expresses all the necessary information to perform predictions.
- The more evidence we get, the more certain we will become of the unknowns, until all but one value combination for the unknowns have probabilities so close to zero that they can be neglected.

C_Example 1: Applying Bayes' Theorem

- Company A is employing workers on short term jobs that are well paid.
- The job sets certain prerequisites to applicants linguistic abilities.
- Earlier all the applicants were interviewed, but nowadays it has become an impossible task as both the number of open vacancies and applicants has increased enormously.
- Personnel department of the company was ordered to develop a questionnaire to preselect the most suitable applicants for the interview.

C_Example 1: Applying Bayes' Theorem

- Psychometrician who developed the instrument estimates that it would work out right on 90 out of 100 applicants, if they are honest.
- We know on the basis of earlier interviews that the terms (linguistic abilities) are valid for one per 100 person living in the target population.
- The question is: If an applicant gets enough points to participate in the interview, is he or she hired for the job (after an interview)?

C_Example 1: Applying Bayes' Theorem

- A priori probability *P*(H) is described by the number of those people in the target population that really are able to meet the requirements of the task (1 out of 100 = .01).
- Counter assumption of the a priori is *P*(~H) that equals to 1–*P*(H), thus it is = .99.
- Psychometricians beliefs about how the instrument works is called conditional probability *P*(E|H) = .9.
- Instruments failure to indicate non-valid applicants, i.e., those that are not able to succeed in the following interview, is stated as *P*(E|~H) that equals to .1.
  - These values need not to sum to one!

C_Example 1: Applying Bayes' Theorem

- A priori probability
- Conditional probability
- Posterior probability

$$P(H|E) = \frac{P(E|H) \cdot P(H)}{P(E|H) \cdot P(H) + P(E|\sim H) \cdot P(\sim H)}$$

$$P(H|E) = \frac{(.9) \bullet (.01)}{(.9) \bullet (.01) + (.1) \bullet (.99)} = .08$$

C_Example 1: Applying Bayes' Theorem

Bayes theorem		
P(H) =	0,01	a priori
P(~H) =	0,99	
P(E|H) =	0,9	conditional probability
P(E|~H) =	0,1	
P(H|E) =	**0,083**	posterior probability

C_Example 1: Applying Bayes' Theorem

- What if the measurement error of the psychometricians instrument would have been 20 per cent?
  - $P(E|H)=0.8 \quad P(E|\sim H)=0.2$

C_Example 1: Applying Bayes' Theorem

Bayes theorem		
P(H) =	0,01	a priori
P(~H) =	0,99	
P(E|H) =	0,8	conditional probability
P(E|~H) =	0,2	
P(H|E) =	**0,039**	posterior probability

C_Example 1: Applying Bayes' Theorem

- What if the measurement error of the psychometricians instrument would have been only one per cent?
  - $P(E|H)=0.99$ $P(E|\sim H)=0.01$

C_Example 1: Applying Bayes' Theorem

Bayes theorem

P(H) =	0,01	a priori
P(~H) =	0,99	
P(E|H) =	0,99	conditional probability
P(E|~H) =	0,01	
P(H|E) =	0,500	posterior probability

C_Example 1: Applying Bayes' Theorem

- Quite often people tend to estimate the probabilities to be too high or low, as they are not able to update their beliefs even in simple decision making tasks when situations change dynamically (Anderson, 1995).

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- One of the most important rules educational science scientific journals apply to judge the scientific merits of any submitted manuscript is that all the reported results should be based on so called 'null hypothesis significance testing procedure' (NHSTP) and its featured product, *p*-value.
- Gigerenzer, Krauss and Vitouch (2004, p. 392) describe 'the null ritual' as follows:
  - 1) Set up a statistical null hypothesis of "no mean difference" or "zero correlation." Don't specify the predictions of your research or of any alternative substantive hypotheses;
  - 2) Use 5 per cent as a convention for rejecting the null. If significant, accept your research hypothesis;
  - 3) Always perform this procedure.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- A *p*–value is the probability of the observed data (or of more extreme data points), given that the null hypothesis H_0 is true, $P(D|H_0)$ (id.).
  - The first common misunderstanding is that the *p*–value of, say *t*–test, would describe how probable it is to have the same result if the study is repeated many times (Thompson, 1994).
  - Gerd Gigerenzer and his colleagues (id., p. 393) call this *replication fallacy* as "$P(D|H_0)$ is confused with $1 - P(D)$."

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- The second misunderstanding, shared by both applied statistics teachers and the students, is that the *p*-value would prove or disprove H_0. However, a significance test can only provide probabilities, not prove or disprove null hypothesis.
- Gigerenzer (id., p. 393) calls this fallacy an *illusion of certainty*: "Despite wishful thinking, $p(D|H_0)$ is not the same as $P(H_0|D)$, and a significance test does not and cannot provide a probability for a hypothesis."
- A Bayesian statistics provide a way of calculating a probability of a hypothesis (discussed later in this section).

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- My statistics course grades (Autumn 2006, $n = 12$) ranged from one to five as follows: 1) $n = 3$; 2) $n = 2$; 3) $n = 4$; 4) $n = 2$; 5) $n = 1$, showing that the lowest grade frequency (”1”) from the course is three (25.0%).
  - Previous data from the same course (2000–2005) shows that only five students out of 107 (4.7%) had the lowest grade.
- Next, I will use the classical statistical approach (the likelihood principle) and Bayesian statistics to calculate if the number of the lowest course grades is exceptionally high on my latest course when compared to my earlier stat courses.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- There are numerous possible reasons behind such development, for example, I have become more critical on my assessment or the students are less motivated in learning quantitative techniques.
- However, I believe that the most important difference between the last and preceding courses is that the assessment was based on a computer exercise with statistical computations.
  - The preceding courses were assessed only with essay answers.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- I assume that the 12 students earned their grade independently (independent observations) of each other as the computer exercise was conducted under my or my assistant's supervision.
- I further assume that the chance of getting the lowest grade (θ), is the same for each student.
  - Therefore X, the number of lowest grades (1) in the scale from 1 to 5 among the 12 students in the latest stat course, has a binomial (12, θ) distribution: $X \sim$ Bin(12, θ).
  - For any integer r between 0 and 12,

$$P(r \mid \theta, n) = \left(\frac{12}{r} \right) \theta^r (1-\theta)^{12-r}$$

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- The expected number of lowest grades is $12(5/107) = 0.561$.
- Theta is obtained by dividing the expected number of lowest grades with the number of students: $0.561 \ / \ 12 \approx 0.05$.
- The null hypothesis is formulated as follows: H_0: $\theta = 0.05$, stating that the rate of the lowest grades from the current stat course is not a big thing and compares to the previous courses rates.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- Three alternative hypotheses are formulated to address the concern of the increased number of lowest grades (6, 7 and 8, respectively): H_1: $\theta = 0.06$; H_2: $\theta = 0.07$; H_3: $\theta = 0.08$.
  - H_1: 12/(107/6) = .67 -> .67/12=.056 ≈ .06
  - H_2: 12/(107/7) = .79 -> .79/12=.065 ≈ .07
  - H_3: 12/(107/8) = .90 -> .90/12=.075 ≈ .08

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- To compare the hypotheses, we calculate binomial distributions for each value of θ.
- For example, the null hypothesis (H_0) calculation yields

$$P(r \mid \theta, n) = \left(\frac{12}{3}\right).05^3(1-.05)^{12-3}$$

$$= \left(\frac{12!}{3!(12-3)!}\right).05^3(1-.05)^{12-3}$$

$$= \left(\frac{479001600}{2177280}\right).05^3(1-.05)^{12-3}$$

$$\approx .017$$

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- The results for the alternative hypotheses are as follows:
  - $P_{H1}(3|.06, 12) \approx .027$;
  - $P_{H2}(3|.07, 12) \approx .039$;
  - $P_{H3}(3|.08, 12) \approx .053$.
- The ratio of the hypotheses is roughly 1:2:2:3 and could be verbally interpreted with statements like “the second and third hypothesis explain the data about equally well”, or “the fourth hypothesis explains the data about three times as well as the first hypothesis”.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- Lavine (1999) reminds that $P(r|\theta, n)$, as a function of r (3) and θ {.05; .06; .07; .08}, describes only how well each hypotheses explains the data; no value of r other than 3 is relevant.
  - For example, $P(4|.05, 12)$ is irrelevant as it does not describe how well any hypothesis explains the data.
  - **This likelihood principle, that is, to base statistical inference only on the observed data and not on a data that might have been observed, is an essential feature of Bayesian approach.**

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- The Fisherian, so called 'classical approach' to test the null hypothesis ($H_0 : \theta = .05$) against the alternative hypothesis ($H_1 : \theta > .05$) is to calculate the *p*–value that defines the probability under H_0 of observing an outcome at least as extreme as the outcome actually observed:

$$p = P(r = 3 | \theta = .05) + P(r = 4 | \theta = .05) + \ldots + P(r = 12 | \theta = .05)$$

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- As an example, the first part of the formula is solved as follows:

$$P(r=3\mid\theta=.05)=\frac{n!}{r!(n-r)!}\theta^{r}(1-\theta)^{n-r}=\frac{12!}{3!(12-3)!}.05^{3}(1-.05)^{12-3}\approx.017$$

n	r	Theta	P(r | Theta, n)
12	3	0,05	0,017331859
12	4	0,05	0,002052457
12	5	0,05	0,000172838
12	6	0,05	0,000010612889170898 4
12	7	0,05	0,000000478776955078125
12	8	0,05	0,000000015749241943359 4
12	9	0,05	0,000000000368403320312 5
12	10	0,05	0,000000000005816894531 25
12	11	0,05	0,000000000000055664062 5
12	12	0,05	0,000000000000002441406 25
		P-value	**0,019568262**

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- After calculations, the *p*–value of .02 would suggest H_0 rejection, if the rejection level of significance is set at 5 per cent.
  - Calculation of *p*–value violates the likelihood principle by using $P(r|\theta, n)$ for values of *r* other than the observed value of $r = 3$ (Lavine, 1999):
    - The summands of $P(4|.05, 12)$, $P(5|.05, 12)$, …, $P(12|.05, 12)$ do not describe how well any hypothesis explains observed data.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- A Bayesian approach will continue from the same point as the classical approach, namely probabilities given by the binomial distributions, but also make use of other relevant sources of *a priori* information.
  - In this domain, it is plausible to think that the computer test ("SPSS exam") would make the number of total failures more probable than in the previous times when the evaluation was based solely on the essays.
  - On the other hand, the computer test has only 40 per cent weight in the equation that defines the final stat course grade: [.3(Essay_1) + .3(Essay_2) + .4(Computer test)]/3 = Final grade.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- Another aspect is to consider the nature of the aforementioned tasks, as the essays are distance work assignments while the computer test is to be performed under observation.
- Perhaps the course grades of my earlier stat courses have a narrower dispersion due to violence of the independent observation assumption?
  - For example, some students may have copy-pasted text from other sources or collaborated without a permission.
- As we see, there are many sources of *a priori* information that I judge to be inconclusive and, thus, define that null hypothesis is as likely to be true or false.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- This a priori judgment is expressed mathematically as $P(H_0) \approx 1/2 \approx P(H_1) + P(H_2) + P(H_3)$.
- I further assume that the alternative hypotheses H_1, H_2 or H_3 share the same likelihood $P(H_1) \approx P(H_2) \approx P(H_3) \approx 1/6$.
- These prior distributions summarize the knowledge about θ prior to incorporating the information from my course grades.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- An application of Bayes' theorem yields

$$P(H_0 \mid r=3) = \frac{P(r=3 \mid H_0)P(H_0)}{P(r=3 \mid H_0)P(H_0)+P(r=3 \mid H_1)P(H_1)+P(r=3 \mid H_2)P(H_2)+P(r=3 \mid H_3)P(H_3)}$$

$$\approx \frac{P(r=3 \mid .017)P(\frac{1}{2})}{P(r=3 \mid .017)P(\frac{1}{2})+P(r=3 \mid .027)P(\frac{1}{6})+P(r=3 \mid .039)P(\frac{1}{6})+P(r=3 \mid .053)P(\frac{1}{6})}$$

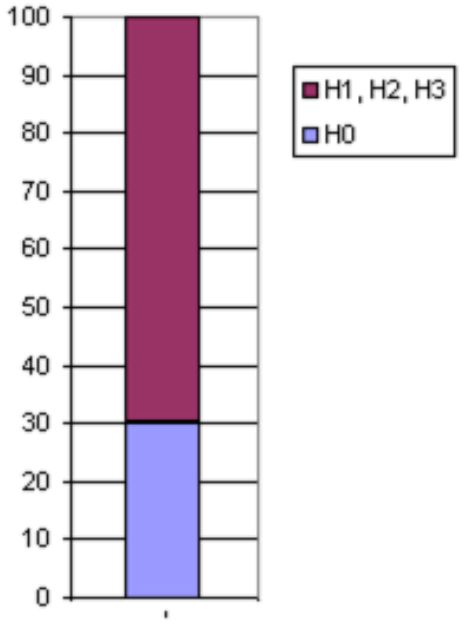
$$\approx 0.30$$

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- Similar calculations for the alternative hypotheses yields $P(H_1|r=3) \approx .16$; $P(H_2|r=3) \approx .29$; $P(H_3|r=3) \approx .31$.
- These *posterior* distributions summarize the knowledge about θ after incorporating the grade information.
- The four hypotheses seem to be about equally likely (.30 vs. .16, .29, .31).
  - The odds are about 2 to 1 (.30 vs. .70) that the latest stat course had higher rate of lowest grades than 0.05.

C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- The difference between the classical and Bayesian statistics would be only philosophical (probability vs. inverse probability) if they would always lead to similar conclusions.
  - In this case the *p*-value would suggest rejection of H_0 ($p = .02$).
  - Bayesian analysis would also suggest evidence against $\theta = .05$ (.30 vs. .70, ratio of .43).



C_Example 2: Comparison of Traditional Frequentistic and Bayesian Approach

- What if the number of the lowest grades in the last course would be two?
  - The classical approach would not anymore suggest H_0 rejection ($p = .12$).
  - Bayesian result would still say that there is more evidence against than for the H_0 (.39 vs. .61, ratio of .64).