Introduction: What is Bayesian and Why

Reminder:

- “applied” : we will require you to use it, so you need to write program codes and we have a midterm on coding.
- “Bayesian”: equations involving Bayes’ rule and integration are inevitable.
- “methods”: algorithms shall be understood and implemented.

Thomas Bayes
(c. 1702 – April 17, 1761)


REV. THOMAS BAYES SON OF THE SAID JOSHUA AND ANN BAYES (59)
7 APRIL 1761.
IN RECOGNITION OF THOMAS BAYES'S IMPORTANT WORK IN PROBABILITY
THIS VAULT WAS RESTORED IN 1969 WITH CONTRIBUTIONS RECEIVED
FROM STATISTICIANS THROUGHOUT THE WORLD
MRS ANN WEST WIDOW OF THE SAID MR THOMAS WES
AND DAUGHTER OF THE SAID JOSHUA AND ANN BAYES
21 DEC. 1758
THEODOSIA BAYES WIFE OF SAMUEL BAYES ESQ (68)
22 SEPT 1769
SAMUEL BAYES ESQ

Outline

- I). Role of statistics
-     with an example of statistical inference
- II). Role of Bayesian statistics
- III). Necessary background
- IV). What is Bayesian
- V). Step1: Get used to Conditional Distribution and play with Bayes' rule

- I). Role of statistics

Deduction and Induction

- Mathematical thinking versus scientific thinking;

Mathematics: deduction

Causes ⇄ Results

Science: induction

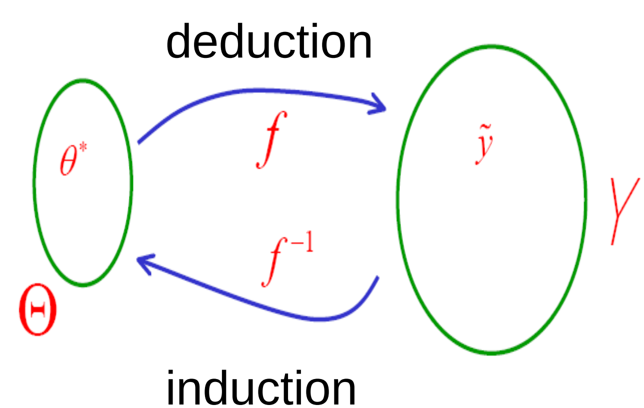
- Probability versus statistics;

Probability: deduction

Conditions	**Probabilistic model**	Observations

Statistics: induction

- Inverse Problem



Statistics in Everything ?

- Quotes:

True logic of this world is in the calculus of probabilities. --- J. C. Maxwell

everybody loves statistics

except mathematicians

But: many mathematicians became statisticians

- Non-experts: stock, gambling, card game, weather, …
- We: statistical inference
- Computer Scientists/Engineers: Data mining, machine learning
- Biologists/chemists/physicists/…: data analysis

  (= hypothesis testing + regression)

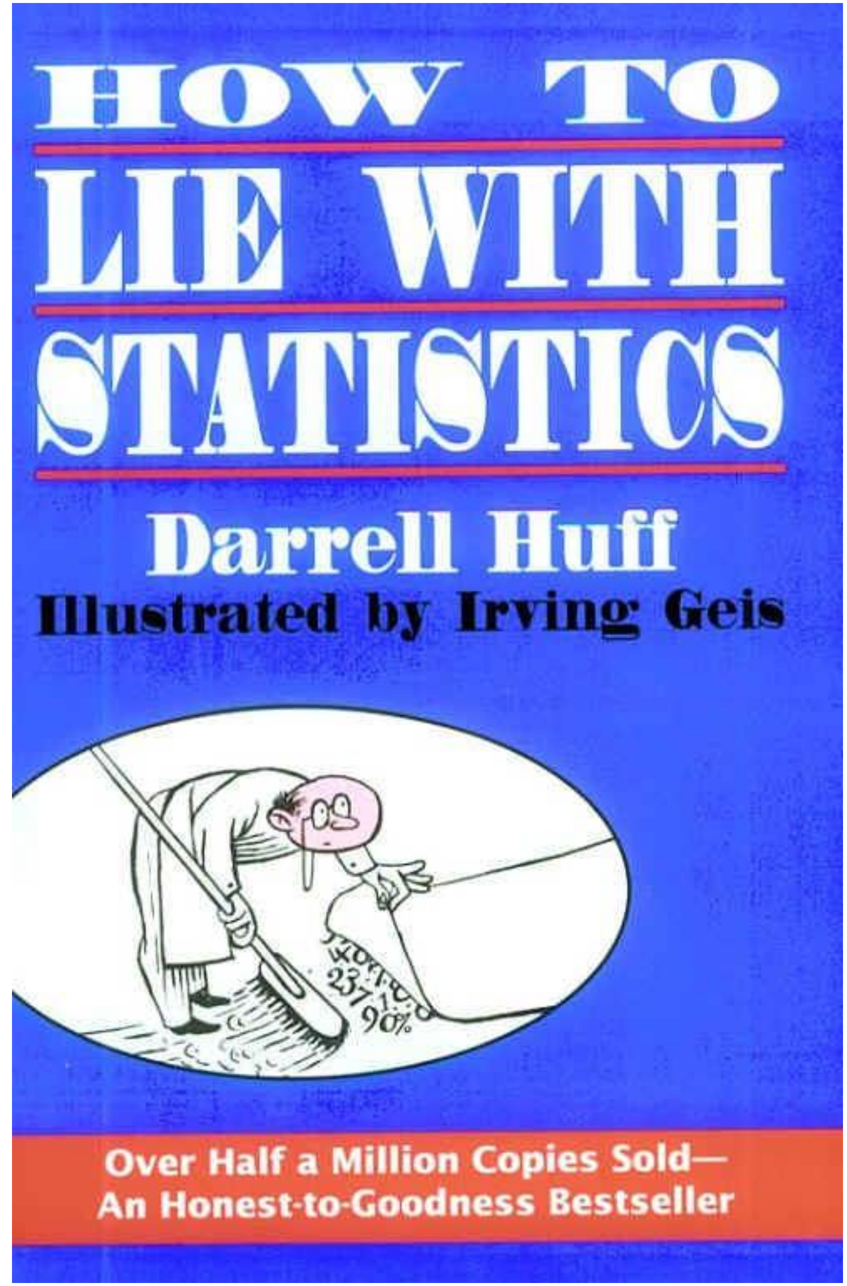
Quotations

- There are three kinds of lies: lies, damn lies, and statistics.

  --Mark Twain =>
  Benjamin Disraeli

- It is easy to lie with statistics, but easier to lie without them.

  --Frederick Mosteller



Role of Probabilistic Models

A device through which we make sense of actual world experiences

- Define primary quantities of interest
- Express scientific knowledge/background
- Describe the data collection process

- Should be reasonably behaved (consistent)
- Should be as simple as possible
- Should be faithful to relevant scientific knowledge (where the randomness come from)

Inference

- To infer: “To conclude based on fact and/or premise”
- Everyday: Make inferences about things unseen based on the observed

Statistical Inference

- Facts are the data
- Premise carried by a probability model (Bayesian: prior distribution and likelihood)
- Conclusions on unknowns (Bayesian: posterior distribution)

Statistical Data Analysis

- A **variable** is a characteristic measured on individuals drawn from a population under study.
- **Data** are measurements of one or more variables made on a collection of individuals.
- Variable/Data Type: discrete, continuous, categorical
- Variable Type: explanatory and response variables

Population ***Sample***

$$
\begin{tikzcd}
\text{probability model for X} \arrow[r, "\text{sample of size n}"] \arrow[d, "\text{deduction}"'] & \text{data } \{x_1, x_2 \ldots, x_n\} \arrow[d, "\text{estimation}"] \\
\text{model mean } \mu_x & \text{sample mean } \hat{\mu}_x \arrow[l, "\text{as n increases}"']
\end{tikzcd}
$$

Parameter ***Estimate***

Common Path of Statistical Studies

- 1. Scientific Goal

e.g. find disease gene, compare efficacy of 2 drugs, …

- 2. Study Design

e.g. observation, experiment, cohort study, case-control, …

- 3. Data Collection

e.g. what variables including confounders to collect, how to collect, …

- 4. Data Describing

e.g. histograms, box plots, average, ...

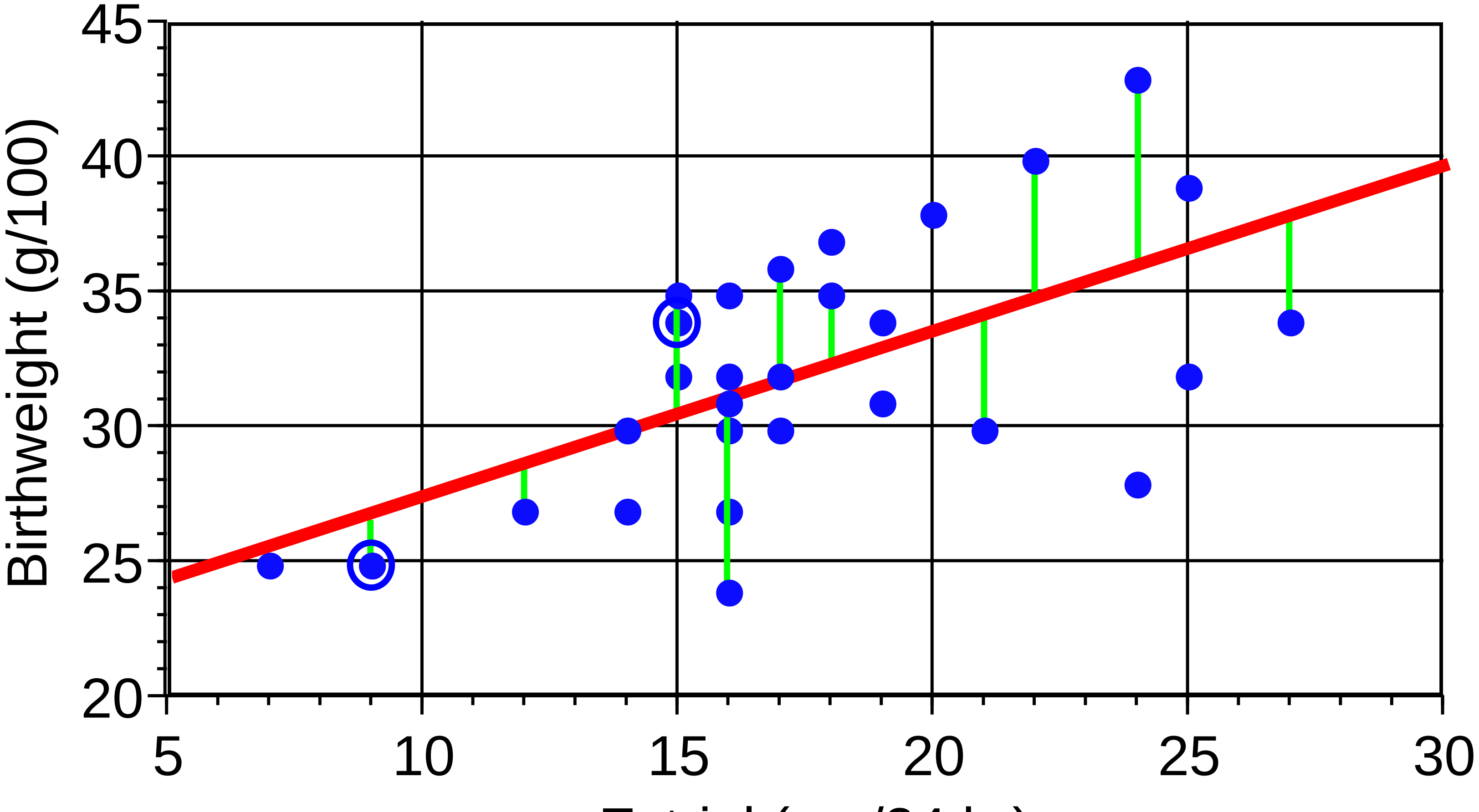
- 5. Statistical Inference

e.g. estimation, hypothesis testing, prediction, …

- 6. Interpretation

- an example of statistical inference

Simple Linear Regression



Rosner Table 11.1

Am J Obs Gyn 1963;85:1-9

Simple Linear Regression

a) The Model

We assume that

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Where x_i is a variable observed on the i^{th} patient

ε_i is assumed to be normally and independently distributed with mean 0 and standard deviation σ

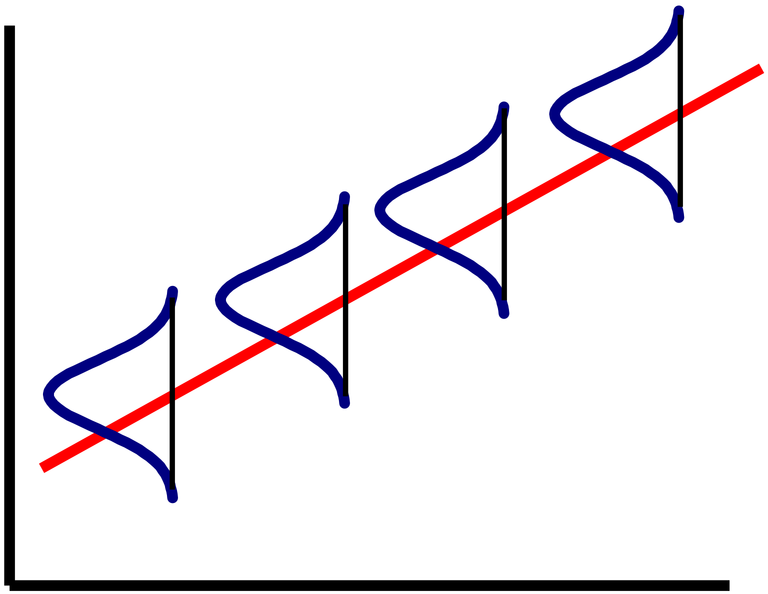
y_i is the response from the i^{th} patient

α and β are model parameters.

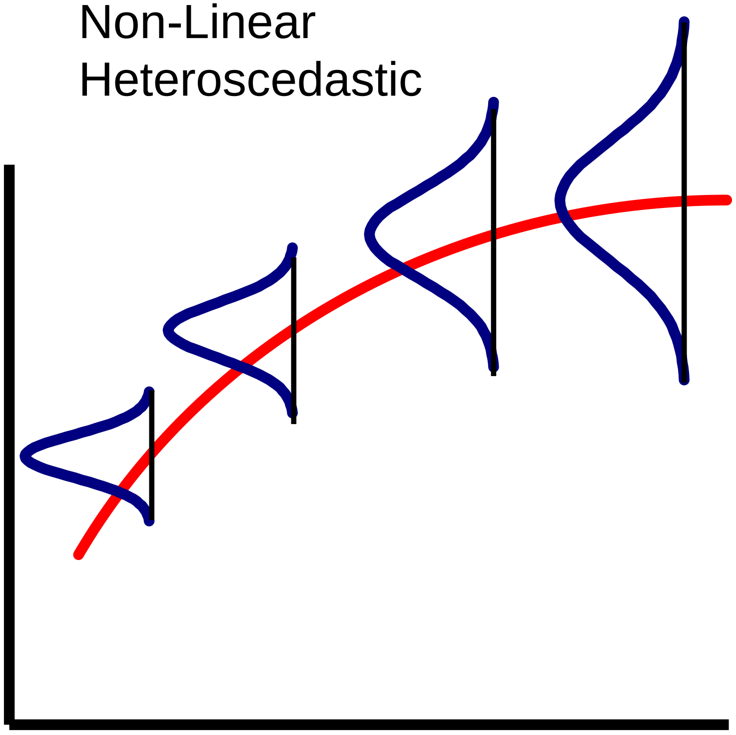
The expected value of y_i is $E(y_i) = \alpha + \beta x_i$.

b) Implications of linearity and homoscedasticity.

Linear
Homoscedastic



Non-Linear
Heteroscedastic

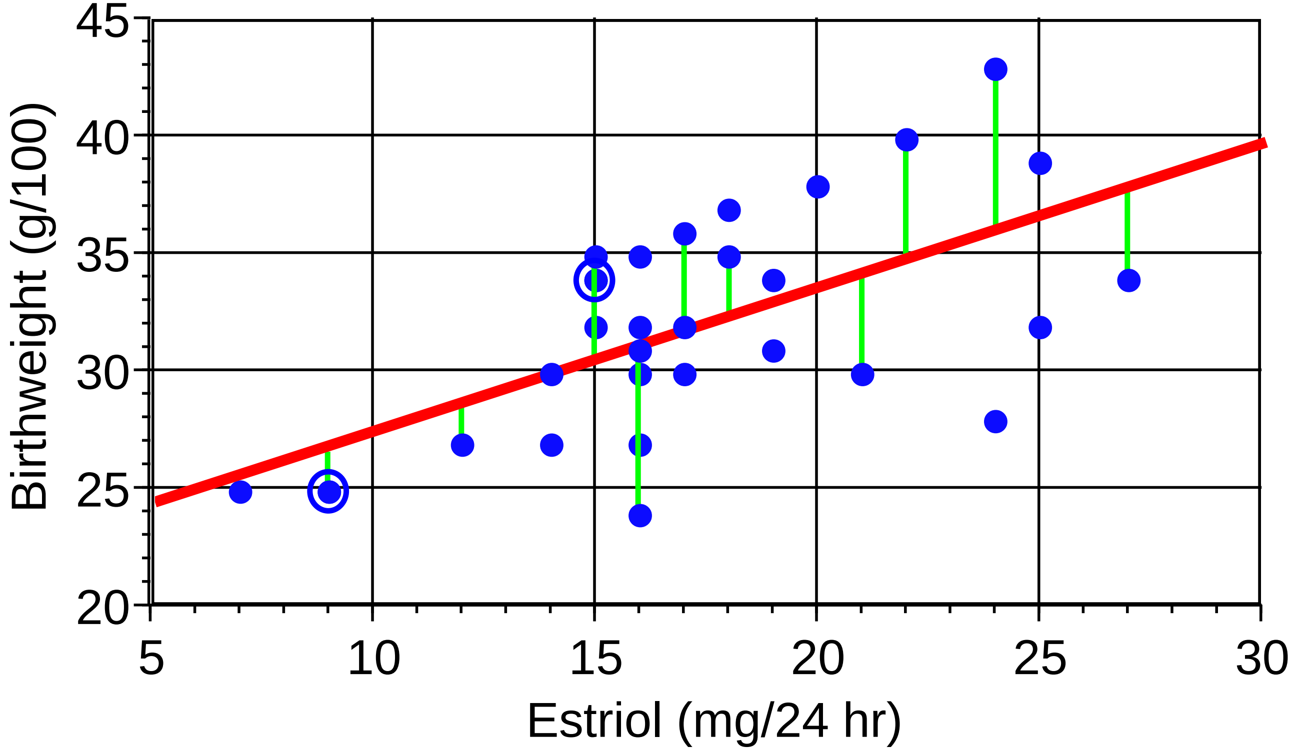


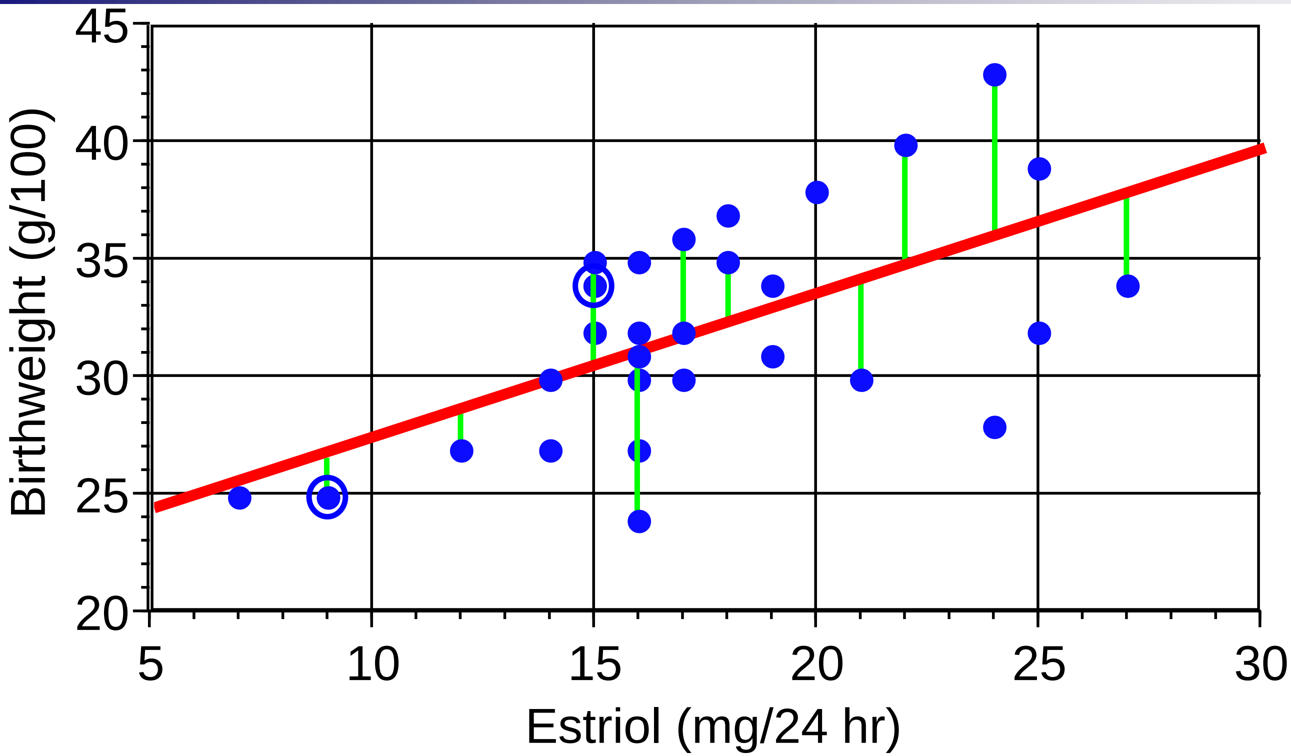
We estimate a and b by minimizing the sum of the squared residuals

$$S=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\sum_{i=1}^{n}\left(y_i-a-bx_i\right)^2$$

*The Least Squares line is the line $y=a+bx$

that minimizes S.



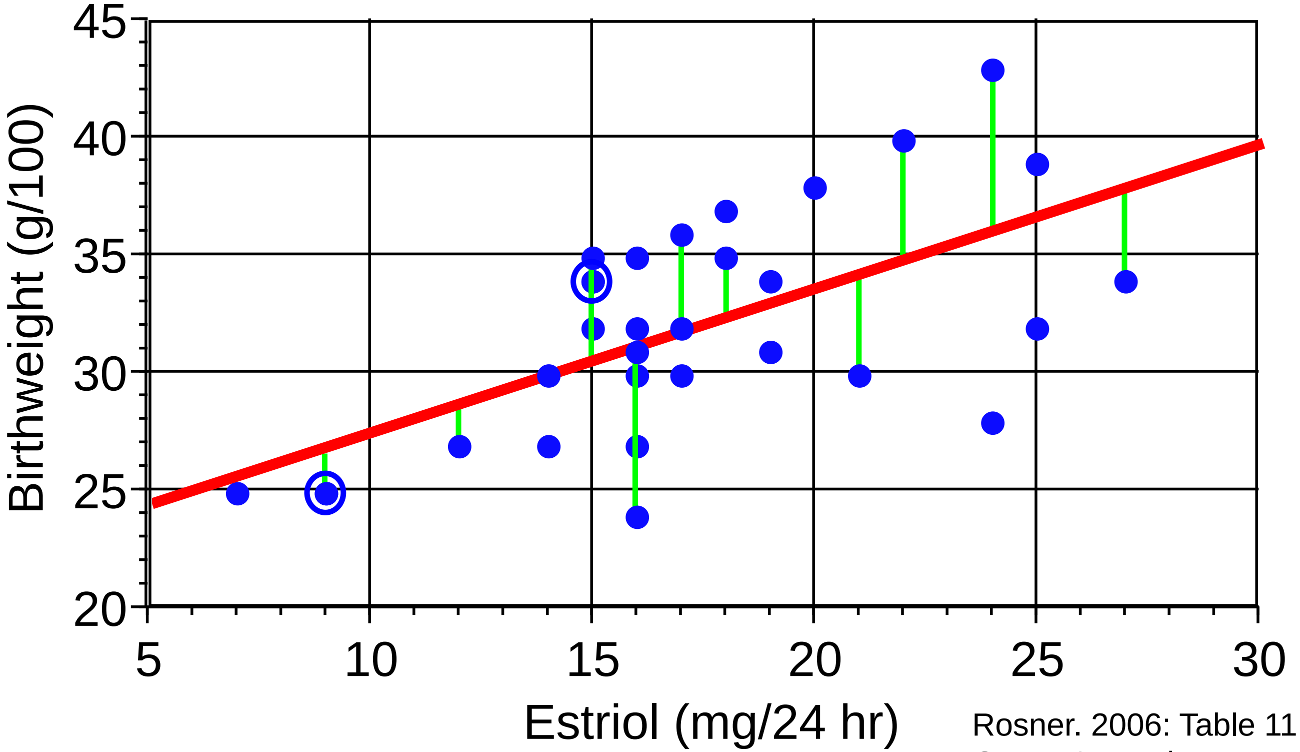


c) Slope parameter estimate

β is estimated by $b = r\, s_y / s_x$

d) Intercept parameter estimate

α is estimated by $a = \bar{y} - b\bar{x}$



Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

e) Least squares estimation

$\hat{y} = a + bx$ is the **least squares estimate** of $\alpha + \beta x$.

II). Role of Bayesian statistics

Methods for inverse problems

Solution procedure

objective

deterministic

- minimum least-squares error
- minimum total absolute error
- minimum maximum absolute error

Non-deterministic

- maximum Likelihood
- maximum a posteriori
- maximum entropy
- minimum mean square error

regularization

deterministic

- function specification (discretization)
- Tikhonov regularization
- iterative regularization
  - --- conjugate gradient
  - --- EM method
- minimum discrepancy principle

- Bayesian prior distribution regularization

optimization

Gradient methods

- Newton's methods
- steepest descent gradient
- conjugate gradient

(sensitivity and/or adjoint problems need to be solved)

Heuristic methods

- greedy search
- simulated annealing
- genetic algorithm
- evolutionary algorithm

- importance/rejection sampling
- MCMC

Bayesian approach

- III). Necessary background

What is Probability? (frequentist)

Consider tossing a coin

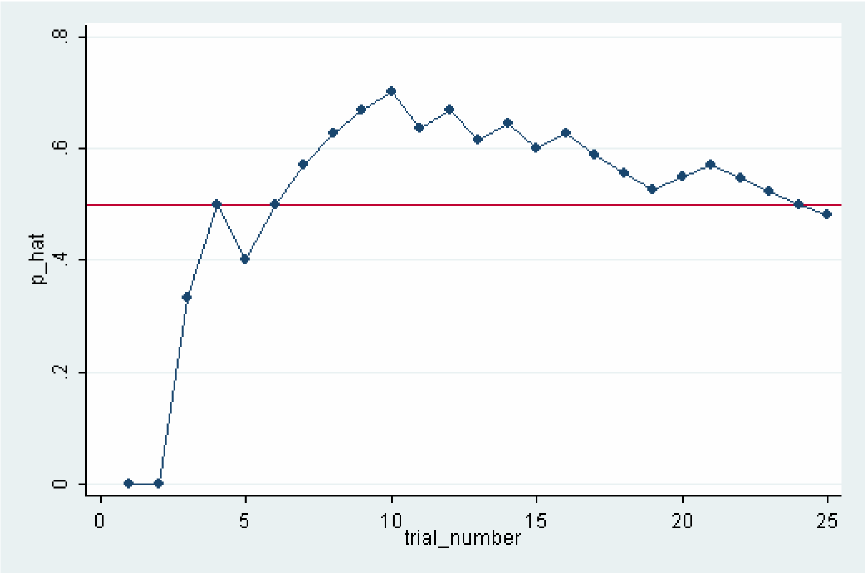
– For any given toss of the coin we can make no nontrivial prediction,

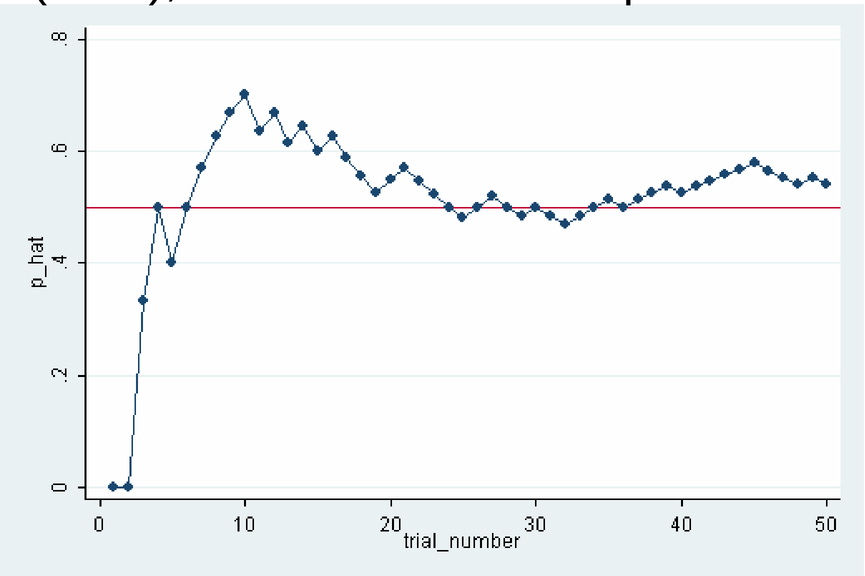
– but observations show that for a large number of tosses the proportion of heads fluctuates around a fixed value, p, between 0 and 1.

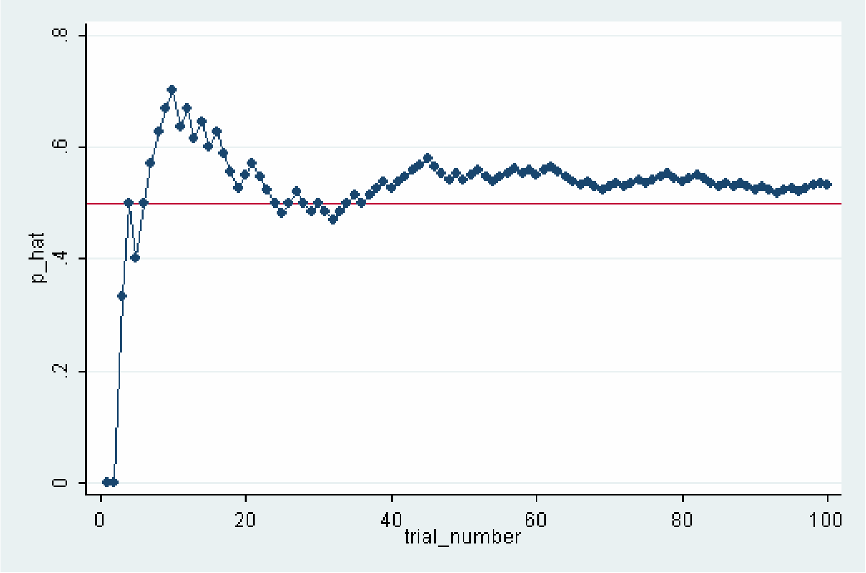
– The proportion of heads appears as it could converge to p, if we let the number of tosses, n, approach infinity. This limiting proportion, p, is called the “probability” that a single toss results in a head.

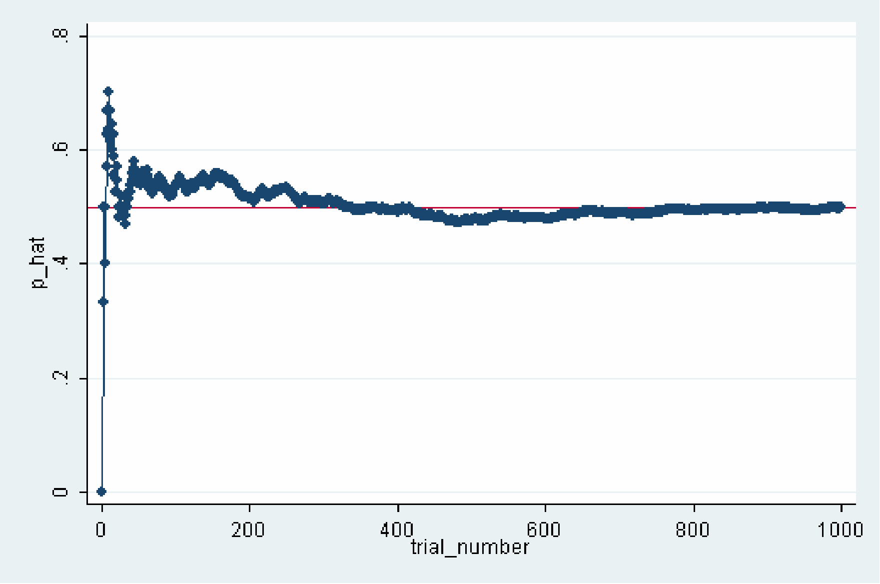
<u>Definition</u>: The probability of an event is the relative frequency of the event over an infinite number of trials.

Experiment for understanding probability: Cumulative estimated probability of head from coin based on 1000 flips (trials), for one series of flips







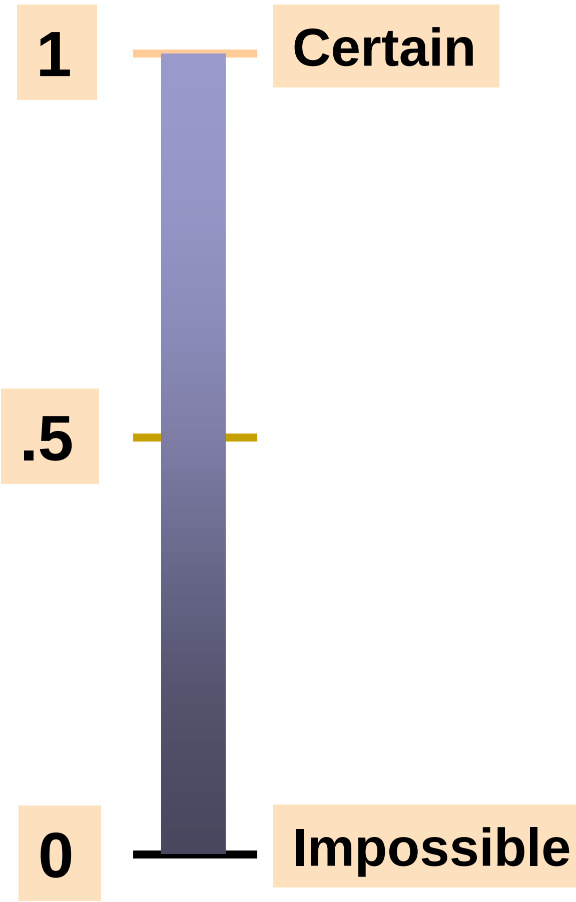


Probability

- Probability – the chance that an uncertain event will occur.

denoted as Pr(E); satisfies $0 \le Pr(E) \le 1$

- e.g., we know that the probability of getting a Head is ½; or Pr{ Head } = ½ and Pr{ Tail } = ½ (p and 1-p for biased coin)
- Other probability of a particular outcome from a random event might be more complicate to imagine, but not unsolvable.
- The probability of getting two heads when tossing a pair of fair coins is ¼



Terminology

- **Population vs. sample**

Population Parameters	**Sample Statistics**
Mean: μ	Average/sample mean: $\bar{x}$
Variance: σ^2	Sample Variance: s^2
Standard Deviation: σ	Sample Standard Deviation: s

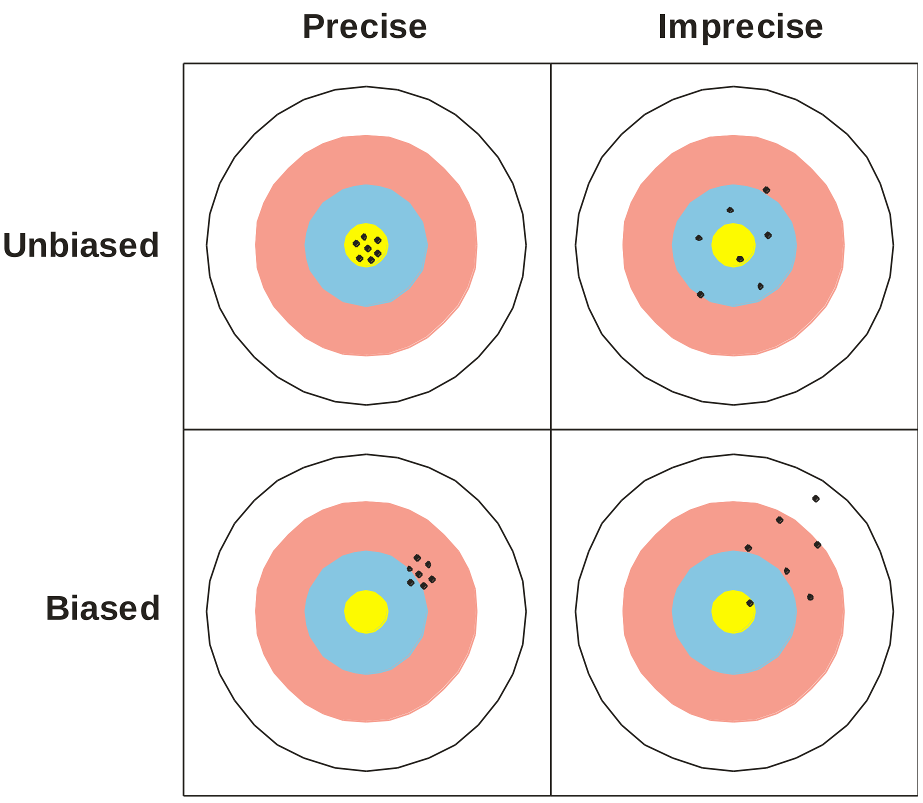
Terminology

•<u>**Properties of a good sample**</u>:
-Independent selection of individuals
-Random selection of individuals
-Sufficiently large

•<u>Independent and identically distributed (i.i.d.) sample:</u> each random variable has the same probability distribution as the others and all are mutually independent

Terminology

- **<u>Bias</u>**<u>:</u> a systematic discrepancy between estimates and the true population characteristic.
- **<u>variability</u>**



Terminology

Estimation is concerned with predicting values or intervals of specific population parameters, based on a set of observed data.

Hypothesis Testing is concerned with testing whether the value of a population parameter is equal to some specific value (or another parameter), based on a set of observed data.

Probability Rules

*Complement Rule: P (A^C) = 1-P (A)

*Addition Rule: if A and B are disjoint events,

then P (A or B) = P (A) + P (B)

*Full Probability Rule: $P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$

*Multiplication Rule: if events A and B are independent, then P (A and B) = P (A)*P (B)

*Bayes' Rule:

$$P(A|B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^c)P(A^c)}$$

mean rule and variance rule

$$\mu_{aX+bY+c} = a\mu_X + b\mu_Y + c$$

$$Var(aX + bY + c) = \sigma^2_{aX+bY+c} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\rho\sigma_X\sigma_Y$$

population correlation coefficient: ρ

(ρ=0 if independent)

Sample correlation coefficient:

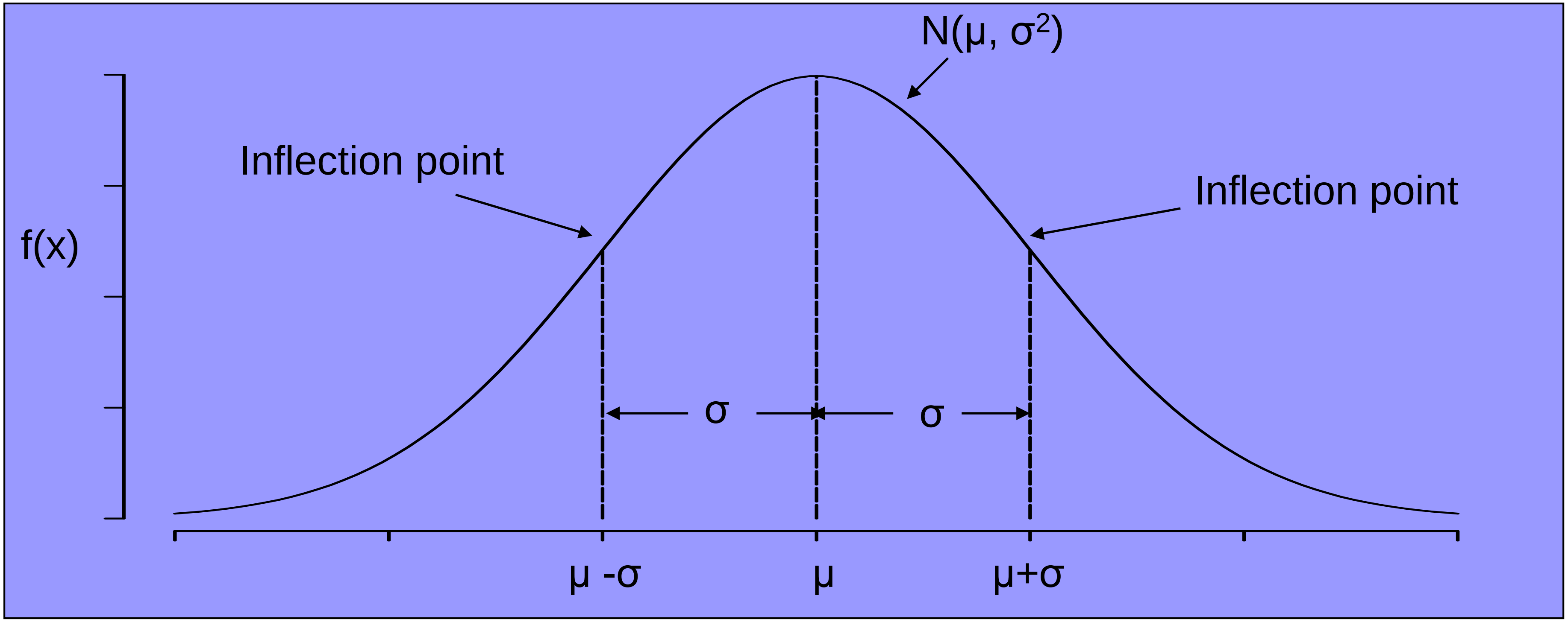
$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

Common Distributions
Normal Distribution

- Probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, \quad -\infty < x < \infty$$

for some parameters μ, σ, where σ>0.



The normal distribution with mean μ and variance σ^2 is denoted by $N(\mu, \sigma^2)$.

Binomial distribution & normal approximation

*X ~ B (n, p),

$$\mu_X = np, \sigma_X^2 = np(1-p)$$

*if X ~ B (n, p) and np>=10, n(1-p)>=10, then approximately X ~ Normal with

$$\mu_X = np$$

$$\sigma_x = \sqrt{np(1-p)}$$

sampling distribution of a sample mean & Central Limit Theorem (CLT)

Draw a SRS of size n from a population which has mean μ

and standard deviation σ

*if the population has normal distribution, then exactly we have:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

*if the population is not normal distribution but n is large, then we still approximately have:

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- IV). What is Bayesian

Frequentists

- The common view of probability is the so-called frequentist approach:

- whereby the probability P of an uncertain event A, P(A), is defined by the frequency of that event based on previous observations.

- For example, in the UK 50.9% of all babies born are girls; suppose then that we are interested in the event A: 'a randomly selected baby is a girl'.

- According to the frequentist approach P(A)=0.509.

Bayesian

- The frequentist approach for defining the probability of an uncertain event is fine providing that we have been able to record accurate information about many past instances of the event. However, if no such historical database exists, then we have to consider a different approach.

- Bayesian probability is a formalism that allows us to reason about beliefs under conditions of uncertainty. If we have observed that a particular event has happened, such as Britain coming 10^{th} in the medal table at the past Olympics, then there is no uncertainty about it.

- However, suppose ***a*** is the statement "Britain sweeps the boards at the next Olympics, winning 36 Gold Medals!"

- Since this is a statement about a future event, nobody can state with any certainty whether or not it is true. Different people may have different beliefs in the statement depending on their specific knowledge of factors that might effect its likelihood.

continued…

- For example, Henry may have a strong belief in the statement ***a*** based on his knowledge of the current team and past achievements.

- Marcel, on the other hand, may have a much weaker belief in the statement based on some inside knowledge about the status of British sport

- Thus, in general, a person's subjective belief in a statement ***a*** will depend on some body of knowledge K. We write this as P(***a***|K). Henry's belief in ***a*** is different from Marcel's because they are using different K's. However, even if they were using the same K they might still have different beliefs in ***a***.

- The expression P(***a***|K) thus represents a *belief measure*. Sometimes, for simplicity, when K remains constant we just write P(***a***), but you must be aware that this is a simplification.

Bayes' Theorem

- True Bayesians actually consider conditional probabilities as more basic than joint probabilities . It is easy to define P(A|B) without reference to the joint probability P(A,B). To see this note that we can rearrange the conditional probability formula to get:

- P(A|B) P(B) = P(A,B)

by symmetry:

- P(B|A) P(A) = P(A,B)

- It follows that:

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

which is the so-called ***Bayes' Rule***

Why are we here?

- Fundamental difference in the experimental approach
- Normally reject (or fail to reject H_0) based on an arbitrarily chosen P value (conventionally <0.05) [In other words we choose our willingness to accept a Type I error]
- This tells us nothing about the probability of H_1
- The frequentist conclusion is restricted to the data at hand, it doesn’t take into account previous, valuable information.

In general, we want to relate an event (E) to a hypothesis (H) and the probability of E given H

The probability of a H being true is determined.

A probability distribution of the parameter or hypothesis is obtained

You can compare the probabilities of different H for a same E

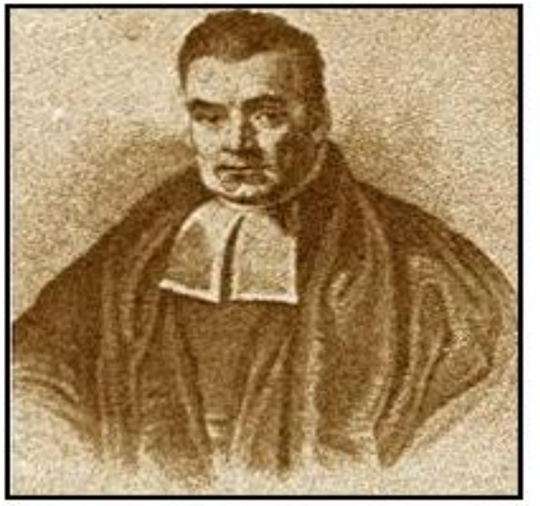
Conclusions depends on previous evidence. Bayesian approach is not data analysis per se, it brings different types of evidence to answer the questions of importance.

Given a prior state of knowledge or belief, it tells how to update beliefs based upon observations (current data).

Fundamentals of Bayesian statistics

- Bayes' formula

$$P(\theta \mid Y) = \frac{P(\mathrm{Y}, \theta)}{P(Y)} = \frac{P(\mathrm{Y} \mid \theta) P(\theta)}{P(Y)}$$



Reverend Thomas Bayes 1702-1761

- Bayesian statistics

prior + evidence => posterior probability

- Bayesian estimation

$$\underbrace{p(\theta \mid Y)}_{\text{Posterior distribution}} \propto \underbrace{p(\theta)}_{\text{Prior distribution}} \underbrace{p(Y \mid \theta)}_{\text{Likelihood}}$$

- A hierarchical formulation

$$\boldsymbol{p(\theta, \lambda \mid Y) \propto p(Y \mid \theta, \lambda) p(\theta \mid \lambda) p(\lambda)}$$

Bayesian inference

- Distinguish between

  Y : known quantities (observed data)

  θ : unknown quantities (e.g. regression coefficients, unknown network structures, missing observations, future outcomes …)

- Fundamental idea: use probability distributions to represent uncertainty about all unknowns
- Likelihood – model for the data: $p(Y | \theta)$
- Prior distribution – represent current uncertainty about unknowns: $p(\theta)$
- Applying Bayes theorem gives posterior distribution:

  prior knowledge + evidence from the data likelihood => posterior probability

$$p(\theta \mid Y) \propto p(\theta)\, p(Y \mid \theta)$$

- Easy to deal with hierarchical structure:

$$\boldsymbol{p(\psi, \lambda \mid Y) \propto p(Y \mid \psi, \lambda)\, p(\psi \mid \lambda)\, p(\lambda)}$$

- Bayesian analyses can be computationally intense, but there have been breakthroughs in algorithms such as the crucial Bayesian tool called Markov Chain Monte Carlo (MCMC).
- Easy to deal with nuisance parameters / missing data: integrate out by MCMC

Philosophy in Bayes' formula

- **Bayes' rule prescribes how learning takes place under uncertainty.**
- **Yesterday's state of knowledge is updated by today's data.**
- **And allows for predicting tomorrow's results.**
- **Formulating what I know today allows for assessing chances of what's going to happen tomorrow.**

Bayesian view: update continually as data accumulate

Overview of Bayesian Approach

- Bayesian methods have been successful in many research areas

  - medicine / epidemiology / genetics
  - ecology / environmental sciences
  - political and social sciences, ………
  - finance
  - archaeology

- Advantage of Bayesian approach
  - natural and coherent way of thinking about science and learning: prior knowledge + evidence from current data => updated/posterior knowledge
  - provides a coherent method for combining evidences: Evidence can accumulate in various ways (sequentially, measurement of many "similar" units, measurement of different aspects of a problem); Evidence can take different forms (data, expert judgment, prior belief)
  - provides a formal framework for fully accounting for uncertainty, propagating uncertainty, and borrow strength to improve precision / effective sample size: well suited to building complex models by linking together multiple sub-models; can obtain estimates and uncertainty intervals for any parameter, function of parameters or predictive quantity of interest
  - exact inference, doesn't rely on asymptotics or analytic approximations: arbitrarily wide range of models can be handled using same inferential framework; focus on specifying realistic models, not on choosing analytically tractable approximation

Bayesian computational framework

- Bayes' formula: $P(\theta \mid Y) = \dfrac{P(Y, \theta)}{P(Y)} = \dfrac{P(Y \mid \theta) P(\theta)}{P(Y)}$

where $P(Y) = \int_{\theta \in \Theta} P(Y, \theta) d\theta = \int_{\theta \in \Theta} P(\theta) \cdot P(Y \mid \theta) d\theta$

Prior distribution modeling
- incorporate prior knowledge
- add physical constraints
- regularize the likelihood

Data Likelihood under the model
- hierarchical structure
- nuisance parameters
- missing data

Hierarchical Bayesian formulation

$$p(\psi, \lambda \mid Y) \propto p(Y \mid \psi, \lambda) p(\psi \mid \lambda) p(\lambda)$$

Posterior distribution exploration (Markov chain Monte Carlo)
- Gibbs sampler
- Metropolis-Hastings sampler
- Reversible Jump MCMC

Posterior distribution can be used as prior distribution for new data

- “If you are not smart enough, be a Bayesian”

Frequentist vs. Bayesian

To a frequentist, unknown model parameters are fixed and only estimable by replications of data from some experiment (repeated trials).

A Bayesian thinks of parameters (at least his knowledge of them) as random, having distributions, like the data (subjective belief).

Bayesian and frequentist probabilities are inverses of each other in the sense that the roles of the arguments and the conditions are reversed.

$P\ (\text{Data} \mid \text{Model}) \neq P\ (\text{Model} \mid \text{Data})$

Model = male or female

Data = pregnant or not pregnant

P (pregnant | female) ~ 3%

but

P (female | pregnant) >>>3%

Frequentist vs. Bayesian

Probability interpretations

- Frequency probability is the interpretation of probability that defines an event's probability as the limit of its relative frequency in a large number of trials.

- Bayesian probability interprets the concept of probability as "a measure of a state of knowledge", a "personal belief".

Bayes versus Frequentism

Both analyse data (x) ➔ statement about parameters (μ)

e.g. Prob ($\mu_l \leq \mu \leq \mu_u$) = 90%

but very different interpretation

Frequentist	μ_l and μ_u known, but random μ unknown, but fixed Probability statement about μ_l and μ_u
Bayesian	μ_l and μ_u known, and fixed μ unknown, and random Probability/credible statement about μ

	Bayesian	Frequentist
Basis of method	Bayes Theorem --> Posterior probability distribution	Uses PDF for data, for fixed parameters
Meaning of probability	Degree of belief	Frequentist defintion
Prob. of parameters?	Yes	Anathema
Needs prior?	Yes	No
Final statement	Posterior probability distribution	Parameter values → Data is likely
Unphysical/ empty range	Excluded by prior	Can occur
Likelihood principle?	Yes	No

V). Step1:
Get used to Conditional Distribution
and
play with Bayes’ rule

Conditional Probability

- Suppose that you know a family has two children, then the probability that both of the children are girls is 1/4. (This is similar of tossing two fair coins and both of them turn up heads)
- If you know in advance that at least one of the children is a girl, then the probability that both of the children are girls is increased to 1/3.
- The later probability is an example of conditional probability.

Conditional Probability

- Conditional probability can be calculated by deleting the sample points which does not fit the condition from the sample space.
- In the unconditional case, the sample space is:

  { (Girl, Girl), (Girl, Boy), (Boy, Girl), (Boy, Boy) }
- In the conditional case, the sample space is:

  { (Girl, Girl), (Girl, Boy), (Boy, Girl)}

Conditional Probability

- Conditional probability may also be calculated from unconditional probability:

$$\Pr(A \mid B) = \frac{\Pr(A \text{ and } B)}{\Pr(B)}$$

- In the last example, *B* is the event that at least one of the children is a girl; *A* is the event that the other child is also a girl. It is easy to see that

$$\Pr(B) = 3/4; \quad \Pr(A \text{ and } B) = 1/4$$

Work on conditional distribution on a table

Proportion of sample of University of Delaware students 1974, N=592. Data adapted from Snee (1974).

joint, marginal and conditional probability

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	.36
Brown	.12	.20	.01	.04	.37
Hazel/ Green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

joint **probabilities: p(e,h)**

For example, p(e=blue,h=black) = .03

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	.36
Brown	.12	.20	.01	.04	.37
Hazel/ Green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

Marginal **probabilities: p(e)**

For example, p(e=blue) = .36

= Σ_h p(e=blue,h)

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	.36
Brown	.12	.20	.01	.04	.37
Hazel/ Green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

***Marginal* probabilities: p(h)**
For example, p(h=black) = .18
$= \Sigma_e$ p(e,h=black)

Marginal probabilities:
p(h) with<u>out</u> info about e

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	.36
Brown	.12	.2[illegible]	.01	.04	.37
Hazel/ Green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

***Conditional* probabilities:**

p(h|e=blue) is p(h) <u>with</u> info that e=blue

$$\text{By definition, } p(h \mid e) = \frac{p(e,h)}{p(e)} \text{ and } p(e \mid h) = \frac{p(e,h)}{p(h)}.$$

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	.36
Brown	.12	.2	.01	.04	.37
Hazel/ Green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

***Conditional* probabilities:**

p(h|e=blue) = p(h,e=blue) / p(e=blue)

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03/.36 =.08	.14/.36 =.39	.16/.36 =.45	.03/.36 =.08	.36/.36 =1

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	.36
Brown	.12	.2[illegible]	.01	.04	.37
Hazel/ Green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

***Conditional* probabilities:**

p(h|e=blue) = p(h,e=blue) / p(e=blue)

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03/.36 =.08	.14/.36 =.39	.16/[illegible] =.45	.03/.36 =.08	.36/.36 =1

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Blue	.03	.14	.16	.03	.36
Brown	.12	.20	.01	.04	.37
Hazel/ Green	.03	.14	.04	.05	.27
	.18	.48	.21	.12	

***Conditional* probabilities:**

p(h|e=brown) = p(h,e=brown) / p(e=brown)

Eye Color	Hair Color				
	Black	Brunette	Blond	Red	
Brown	.12/.37 =.32	.20/.37 =.54	.01/.37 =.03	.04/.37 =.11	.37/.37 =1

Bayes' rule

By definition, $p(h \mid e) = \dfrac{p(e,h)}{p(e)}$ and $p(e \mid h) = \dfrac{p(e,h)}{p(h)}$

Hence $p(h \mid e)p(e) = p(e,h) = p(e \mid h)p(h)$, and

$$p(h \mid e) = p(e \mid h)\, p(h) \,/\, p(e)$$

<u>with</u> info about *e*

with<u>out</u> info about *e*

Case Study: An Example of Conditional Probability

- Everyone gets sick sometimes and has to do some medical tests
- Suppose some typical medical test for a rare but potentially fatal disease (say HIV) has a false positive probability of 1 in 100 and true positive probability of 90 in 100.
- By false positive probability we mean the probability that a healthy person come out of the test with a positive result (incorrect result)

Case Study: An Example of Conditional Probability

- By true positive probability we mean the probability that a sick person come out of the test with a positive result (Correct result)
- Suppose that we know for a typical person like me, the probability of having such a disease is small, say, 1 in 1000. (So the probability of no such disease is 999/1000)
- Suppose I go through such a test in a routine medical check-up, and the result is positive, should I be very much concerned?

Case Study: An Example of Conditional Probability

- Since the false positive probability is 1/100, is it correct to say that, since I am tested positive, I have a 99/100 probability or chance to have this disease?
- Do you agree with the following argument?

if <u>I am not sick</u>, then there is only 1/100 chance that the test is incorrect. Since there is 99/100 chance the test is correct, so there is 99/100 chance that <u>I am sick.</u>

- Do you see the problem in this logic? (From the assumption that <u>I am not sick</u> , the conclusion is with 99 out of 100 that <u>I am sick.</u>)

This is the kind of mistake most of us make everyday!

It's hard to get it right the first time!

Bayes' Rule

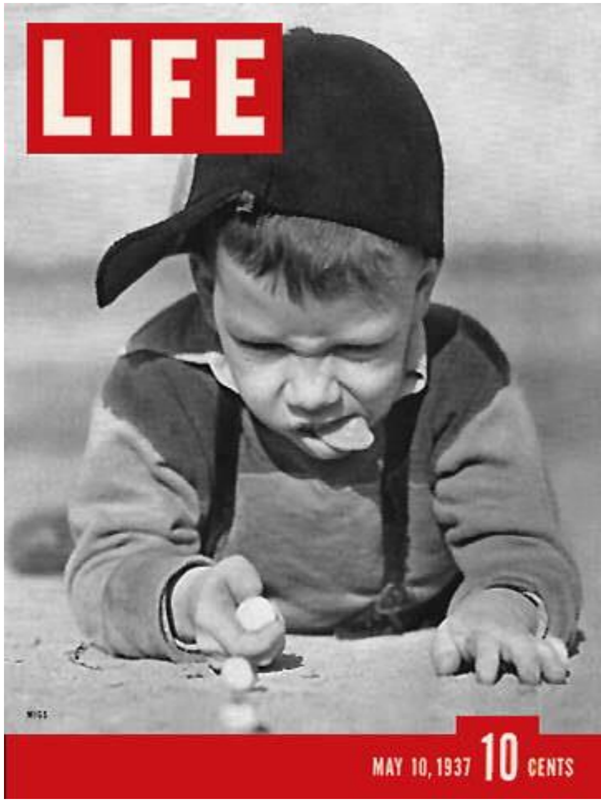
- Let's use some simple mathematical notation for the quantity discussed:
- False positive probability:

  Pr{ Test Positive | Healthy} = 1/100
- True positive probability:

  Pr{ Test Positive | Sick} = 90/100
- Disease prevalence:

  Pr{Healthy} = 999/1000

  Pr{Sick} = 1/1000



Bayes’ Rule

- What is the most important probability after I tested positive is not 1 minus the false positive, it is rather the probability that I have the disease conditional on the fact that I tested positive, or:

  $$\Pr\{\text{ Sick } | \text{ Test Positive } \}$$

- This probability can be calculated using the Bayes’ Rule

Bayes' Rule

$$\Pr\{\text{ Sick} \mid \text{Test Positive }\}$$

$$= \frac{\Pr\{\text{ Sick and Test Positive }\}}{\Pr\{\text{ Test Positive }\}}$$

The numerator can be calculated by the chain rule:

$$\Pr\{\text{ Sick and Test Positive }\}$$

$$= \Pr\{\text{ Sick }\}\Pr\{\text{ Test Positive} \mid \text{Sick }\}$$

$$= \frac{1}{1000} * \frac{90}{100} = \frac{90}{100{,}000}$$

Bayes' Rule

The denominator can be calculated by

$$\Pr\{\text{ Test Positive }\} = \Pr\{\text{ Sick and Test Positive }\} + \Pr\{\text{ Healthy and Test Positive }\}$$

$$= \Pr\{\text{ Sick}\}\Pr\{\text{ Test Positive }|\text{Sick}\} + \Pr\{\text{ Healthy}\}\Pr\{\text{ Test Positive }|\text{Healthy}\}$$

$$= \frac{1}{1000} * \frac{90}{100} + \frac{999}{1000} * \frac{1}{100} = \frac{90+999}{100,000}$$

Bayes' Rule

- Finally Pr{ Sick Person | Test Positive } is about 90/(90+999) which is less that 1/12.
- So the probability that I have the disease after I tested positive is not 99 out of 100 (which is almost certain), but smaller than 1 out of 12! This chance is still quite small.

Understanding Bayes' Rule

- Why the probability is still quite small?
- Suppose that 10,000 people go through this test, out of which 10 persons are sick and 9,990 are healthy. Since the false positive probability is 1/100, about 100 healthy persons will tested positive with the 9 sick persons who tested positive. So out of 109 persons who tested positive, only 9 are actually sick, a rate of 1 in 12.

Bayes' rule

$$p(h \mid e) = p(e \mid h)\, p(h) \,/\, p(e)$$

So what's the big deal?
→ Ramifications when applied to data and parameter values!

$$\underbrace{p(\theta \mid D)}_{\text{posterior}} = \underbrace{p(D \mid \theta)}_{\text{likelihood}} \; \underbrace{p(\theta)}_{\text{prior}} \,/\, \underbrace{p(D)}_{\text{evidence}}$$

posterior ↑ **<u>with</u> info about *D***

prior ↑ **with<u>out</u> info about *D***

Bayes' rule applied to data and parameter values

Data	Parameter			
	…	θ	…	
⋮		⋮		⋮
D	…	$p(D,\theta)=p(D\mid\theta)p(\theta)$	…	$p(D)$
⋮		⋮		⋮
	…	$p(\theta)$	…	

Possible values of parameter

Possible values of data

Bayes' rule applied to data and parameter values

Data	Parameter			
	…	θ	…	
⋮		⋮		⋮
D	…	$p(D,\theta)=p(D\mid\theta)p(\theta)$	…	$p(D)$
⋮		⋮		⋮
	…	$p(\theta)$	…	

Marginal probabilities:
$p(\theta)$ with<u>out</u> info about D
a.k.a. "prior"

Bayes' rule applied to data and parameter values

Data	Parameter			
	...	θ	...	
⋮		⋮		⋮
D	...	$p(D,\theta)=p(D\|\theta)p(\theta)$	...	$p(D)$
⋮		⋮		⋮
	...	$p(\theta)$	...	

Conditional probabilities:

$p(\theta|D)$ is $p(\theta)$ <u>with</u> info about D

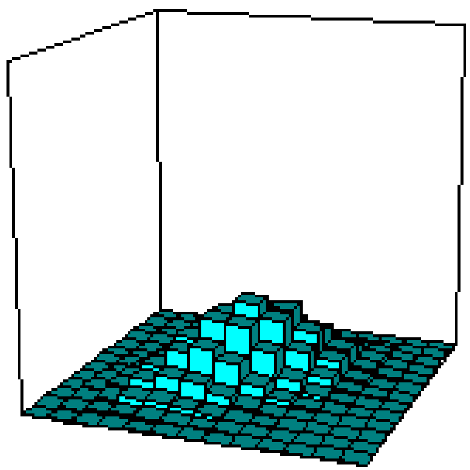
a.k.a. "posterior"

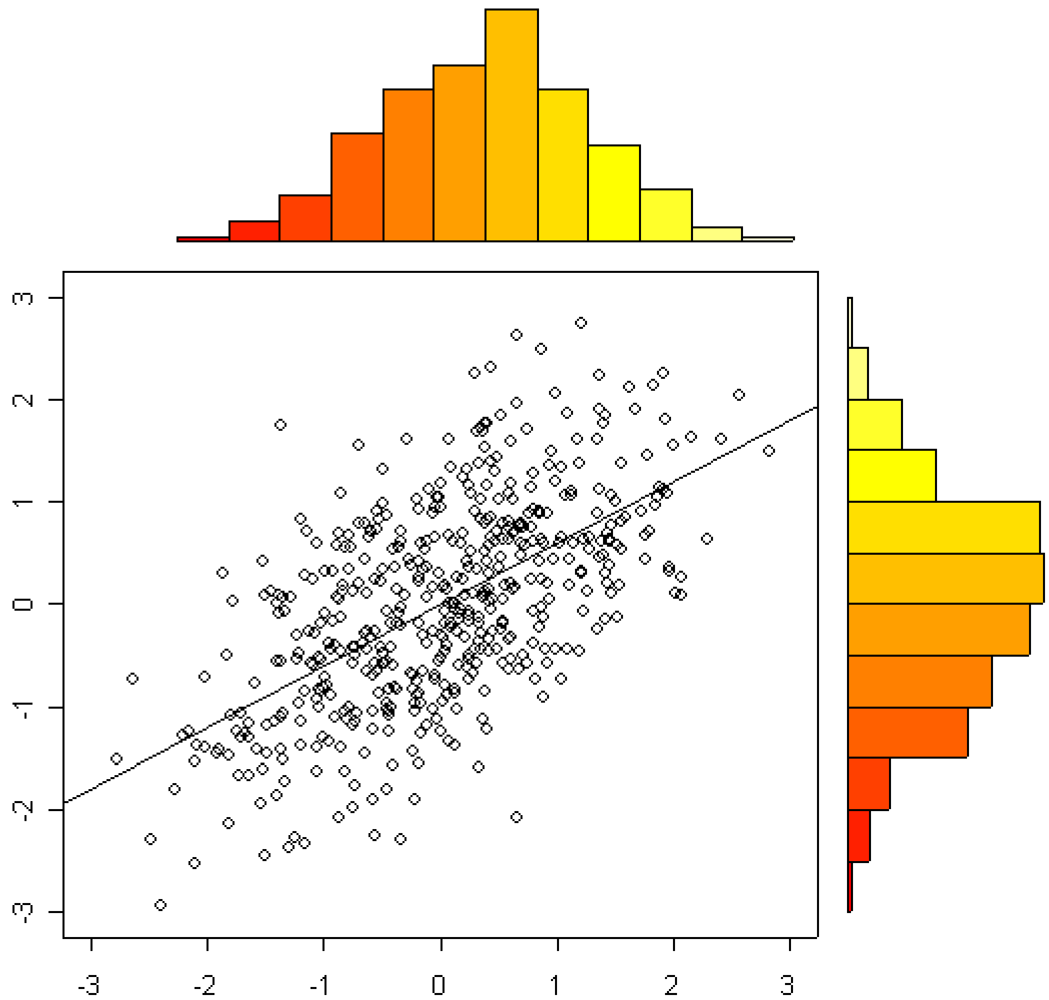
Multi-Parameter Models: Joint, Marginal

- Two random variables: $\theta 1, \theta 2$
- Joint distribution: $P(\theta 1,\theta 2)$
- Marginal distribution: $p(\theta 1)$, $p(\theta 2)$

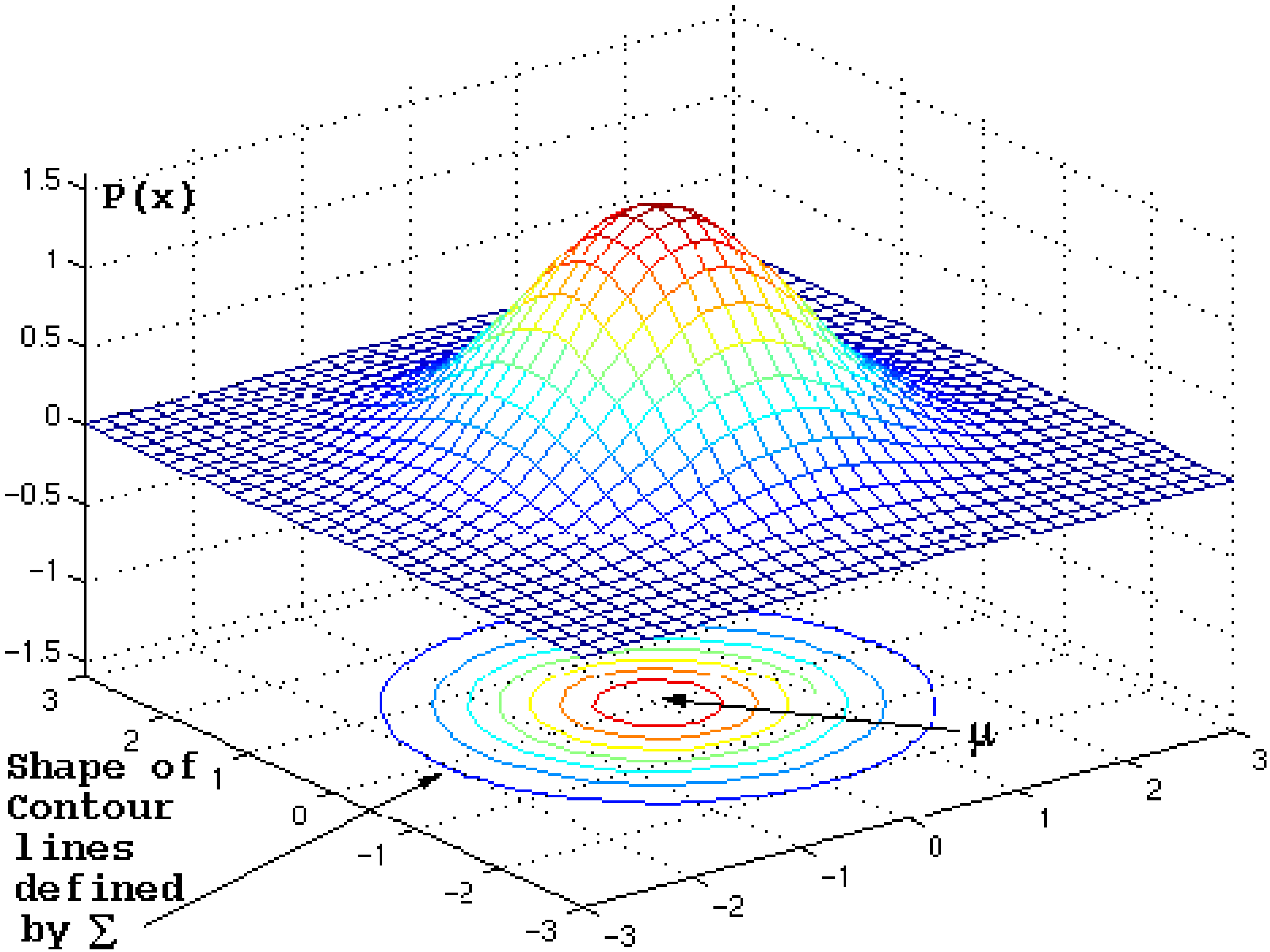
$P(\theta 1)=\int p(\theta 1,\theta 2)d\theta 2$

$P(\theta 2)=\int p(\theta 1,\theta 2)d\theta 1$



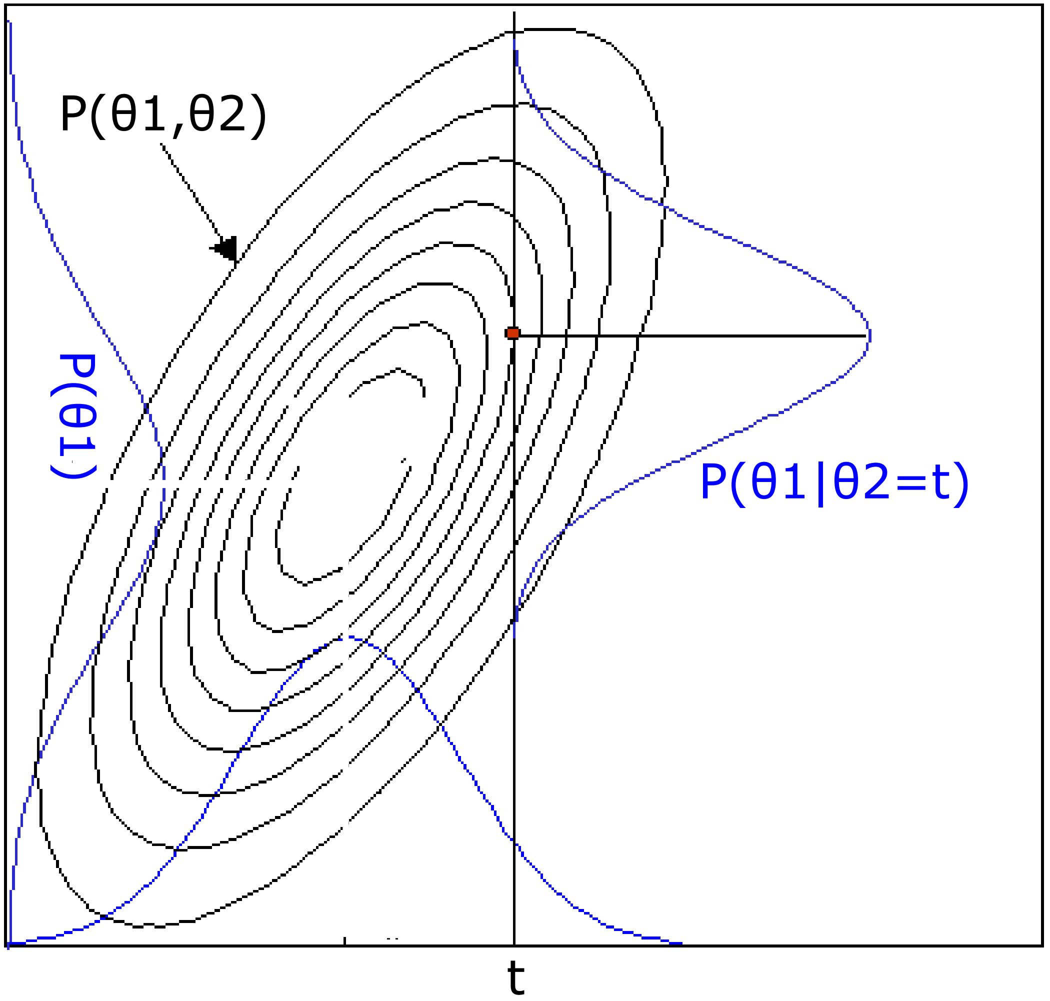


3
2
1
0
-1
-2
-3
-3 -2 -1 0 1 2 3

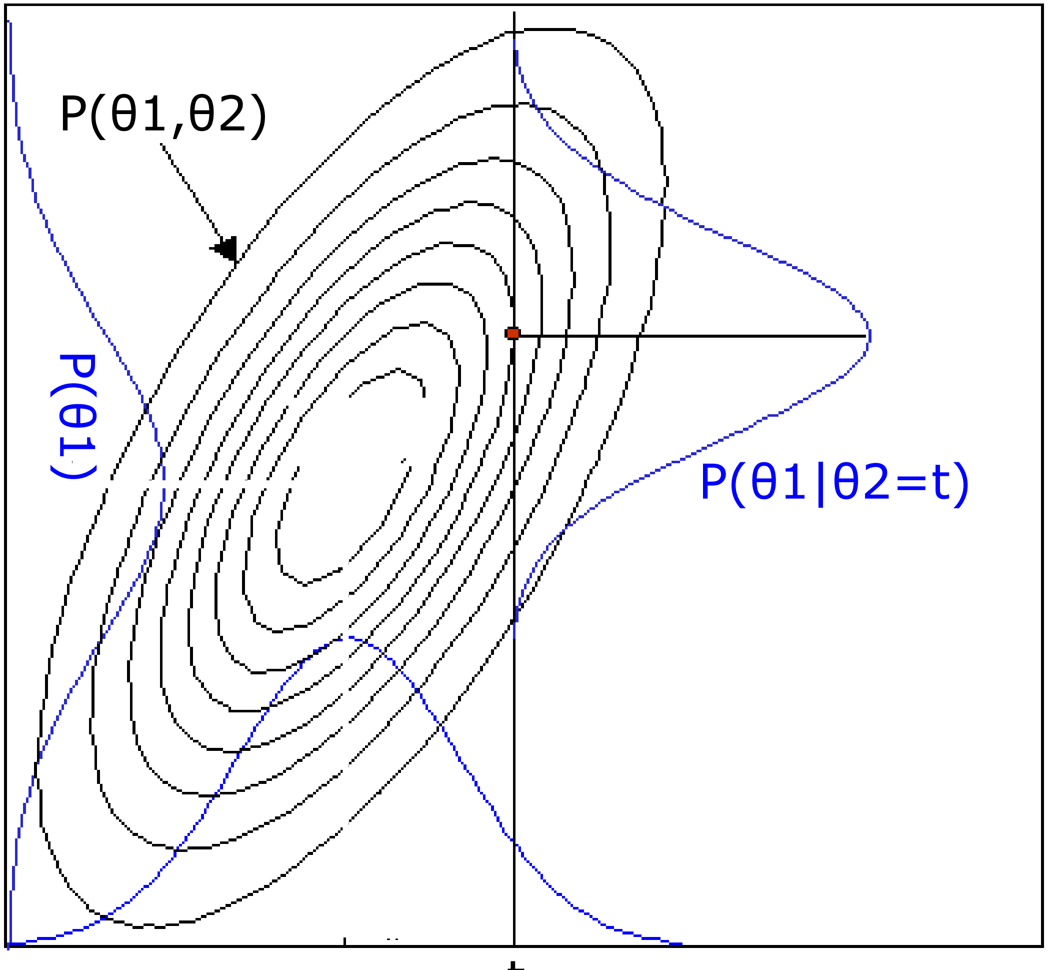


Contour definition: www.byclb.com/TR/Tutorials/neural_networks/ch4_1.htm

Joint, Marginal, Conditional



Joint, Marginal, Conditional



P(θ1,θ2)=p(θ1|θ2)p(θ2)