Lecture 3:

Maximum Likelihood Estimation (MLE), Linear Regression, Linear Models

Markus Hohle

University California, Berkeley

Bayesian Data Analysis and Machine Learning for Physical Sciences

Bayesian Data Analysis and Machine Learning for Physical Sciences

Course Map

Module 1	Maximum Entropy and Information, Bayes Theorem
Module 2	Naive Bayes, Bayesian Parameter Estimation, MAP
Module 3	**MLE, Lin Regression, Model selection: Comparing Distributions**
Module 4	Model Selection: Bayesian Signal Detection
Module 5	Variational Bayes, Expectation Maximization
Module 6	Stochastic Processes
Module 7	Monte Carlo Methods
Module 8	Markov Models, Graphs
Module 9	Machine Learning Overview, Supervised Methods
Module 10	Unsupervised Methods
Module 11	ANN: Perceptron, Backpropagation
Module 12	ANN: Basic Architecture, Regression vs Classification, Backpropagation again
Module 13	Convolution and Image Classification and Segmentation
Module 14	TBD (GNNs)
Module 15	TBD (RNNs and LSTMs)
Module 16	TBD (Transformer and LLMs)

Bayesian Model Selection/Testing

Outline

Standard Tests

- the p-value

Bayesian Model Testing

- idea
- curve fitting revisited
- comparing distributions

Bayesian Model Selection/Testing

Outline

Standard Tests

- the p-value

Bayesian Model Testing

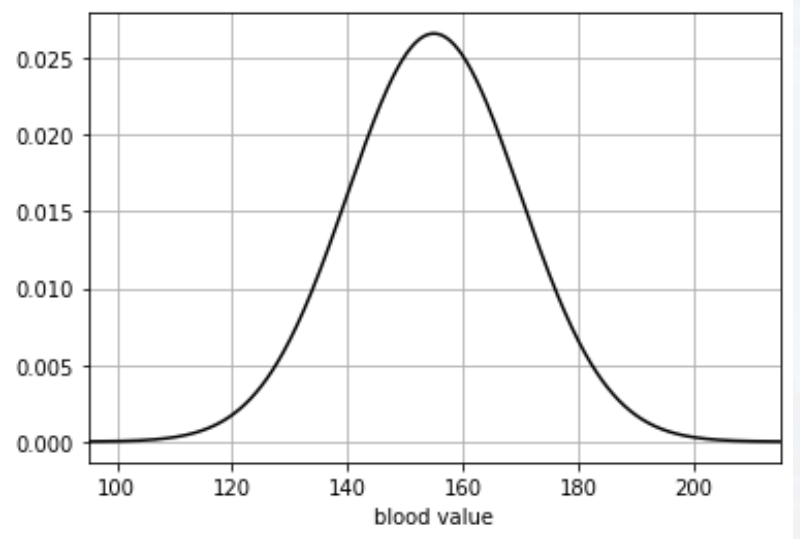
- idea
- curve fitting revisited
- comparing distributions

Bayesian Model Selection/Testing: the p-value

standard frequentist way (not Bayesian):

1) assume a **likelihood function *L* as your model**, aka **null hypothesis H_0**
2) take a datapoint x_0
3) calculate the ***probability P*** given ***L*** i. e. **given H_0** is true, **that x_0 has this value** or a **more extreme value**
4) accepting or rejecting H_0 based on ***P*** and the threshold α

example: 1) a healthy person has a blood value that follows a **normal distribution** with $\mu = 155, \sigma = 15$, i. e. $H_0: N(\mu = 155, \sigma = 15)$



Bayesian Model Selection/Testing: the p-value

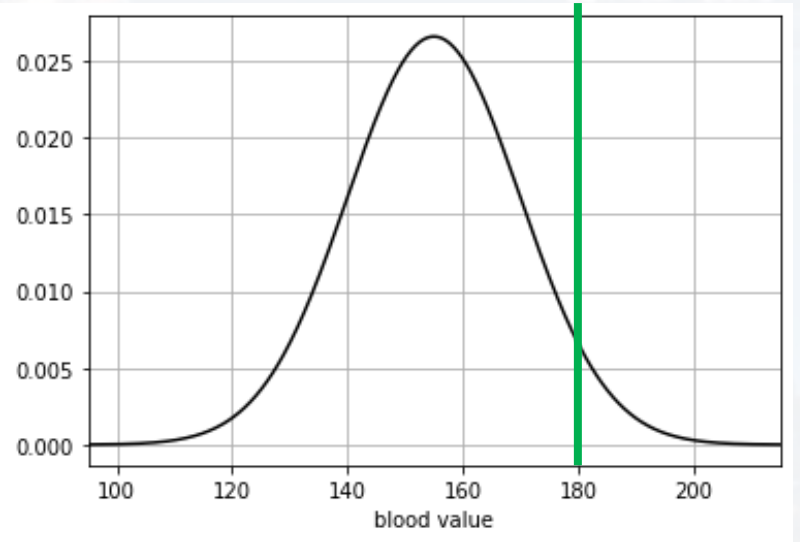
standard frequentist way (not Bayesian):

1) assume a **likelihood function *L* as your model**, aka **null hypothesis H_0**
2) take a datapoint x_0
3) calculate the ***probability P*** given ***L*** i. e. **given H_0** is true, **that x_0 has this value** or a **more extreme value**
4) accepting or rejecting H_0 based on ***P*** and the threshold α

example:

1) a healthy person has a blood value that follows a **normal distribution** with $\mu = 155, \sigma = 15$, i. e. $H_0: N(\mu = 155, \sigma = 15)$

2) a patient has the value x_0 = 180



Bayesian Model Selection/Testing:

standard frequentist way (not Bayesian):

1) assume a **likelihood function *L* as your model**, aka **null hypothesis H_0**
2) take a datapoint x_0
3) calculate the ***probability P*** given ***L*** i. e. **given H_0** is true, **that x_0 has this value** or a **more extreme value**
4) accepting or rejecting H_0 based on ***P*** and the threshold α

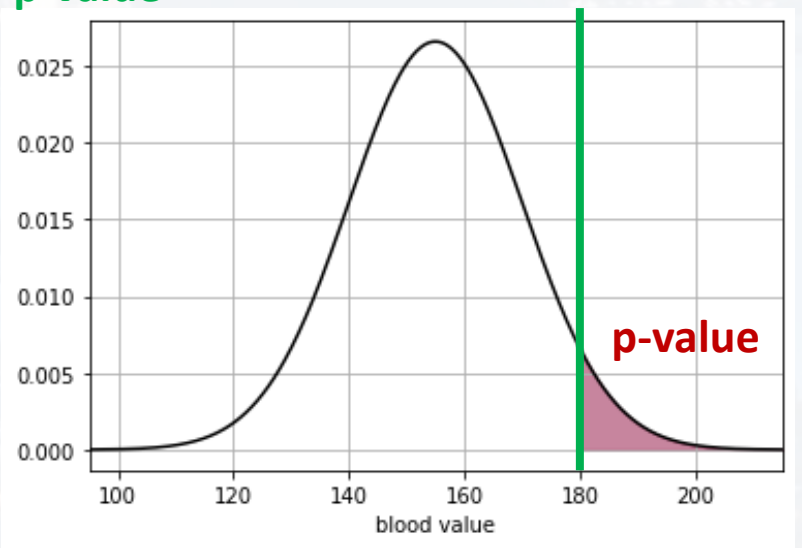
example:

1) a healthy person has a blood value that follows a **normal distribution** with

$\mu = 155, \sigma = 15$, i. e. $H_0: N(\mu = 155, \sigma = 15)$

2) a patient has the value $x_0 = 180$

3) probability $P(x \geq 180 | H_0) = 0.048$ called **p-value**



standard frequentist way (not Bayesian):

1) assume a **likelihood function *L* as your model**, aka **null hypothesis H_0**
2) take a datapoint x_0
3) calculate the ***probability P*** given ***L*** i. e. **given H_0** is true, **that x_0 has this value** or a **more extreme value**
4) accepting or rejecting H_0 based on ***P*** and the threshold α

example:

1) a healthy person has a blood value that follows a **normal distribution** with

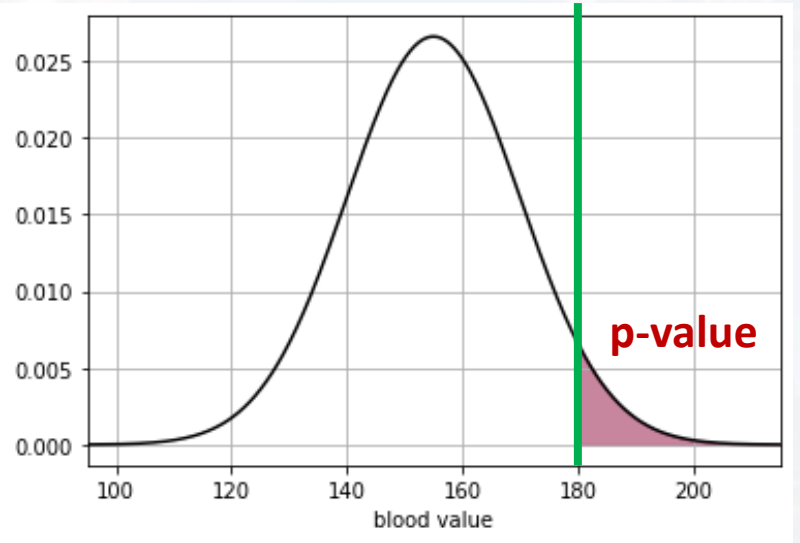
$\mu = 155, \sigma = 15$, i. e. $H_0: N(\mu = 155, \sigma = 15)$

2) a patient has the value $x_0 = 180$

3) probability $P(x \geq 180 | H_0) = 0.048$ called **p-value**

4) for **α = 0.05**, H_0 is **rejected**, i. e. patient is **not healthy**

→ alternative hypothesis H_1: **not healthy**



Bayesian Model Selection/Testing:

example:

1) a healthy person has a blood value that follows a **normal distribution** with

$\mu = 155, \sigma = 15$, i. e. $\boldsymbol{H_0: N(\mu = 155, \sigma = 15)}$

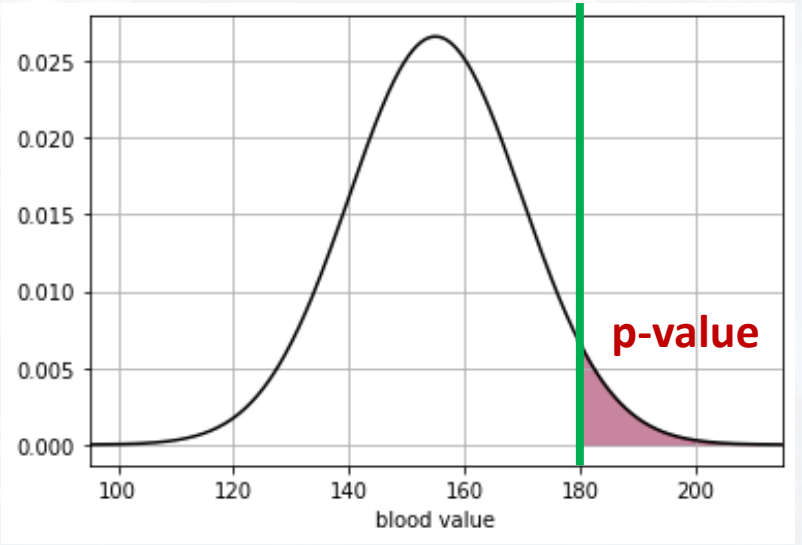
2) a patient has the value $\boldsymbol{x_0 = 180}$

3) probability $\boldsymbol{P(x \geq 180 | H_0) = 0.048}$ called **p-value**

4) for **α = 0.05**, H_0 is **rejected**, i. e. patient is **not healthy**

→ alternative hypothesis $\boldsymbol{H_1}$: **not healthy**

We just performed a so-called Z – Test
(comparing one value to a normal distribution)



Bayesian Model Selection/Testing:

standard frequentist way (not Bayesian):

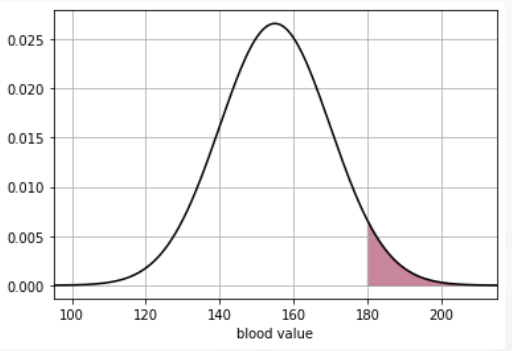
1) assume a **likelihood function *L* as your model**, aka **null hypothesis H_0**
2) take a datapoint x_0
3) calculate the ***probability P*** given ***L*** i. e. **given H_0** is true, **that x_0 has this value** or a **more extreme value**
4) accepting or rejecting H_0 based on ***P*** and the threshold α

caveats:

- the model ***L*** for H_0 has to **be known**
- the p-value $P(x \geq x_0|H_0)$ **does not** tell if H_0 is true or not
- the p-value $P(x \geq x_0|H_0)$ **does not** tell which hypothesis is more likely
- the p-value just gives $P(x \geq x_0|H_0)$
- the threshold α for accepting/rejecting H_0 is **arbitrary**
- we are aiming on disproving a hypothesis, by assuming it is true, **without** leading to a contradiction

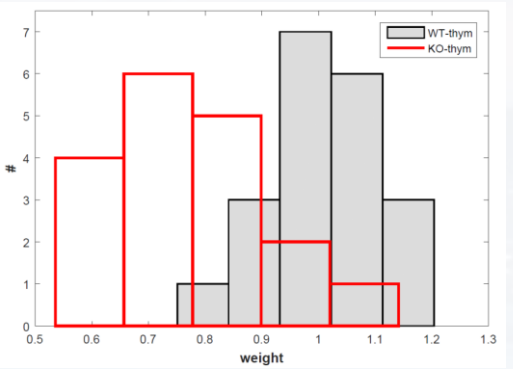
Bayesian Model Selection/Testing:

- data point *x* versus **normal** — Z-test
- $H_0: x \in N(\mu, \sigma)$



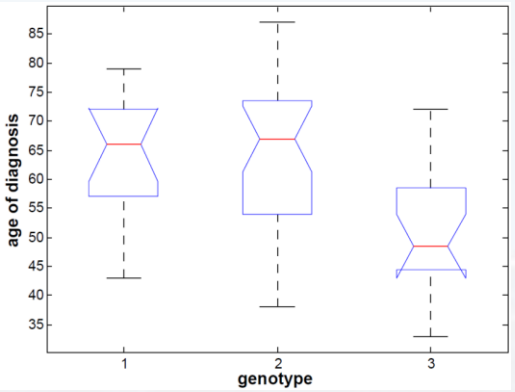
- **normal** versus another **normal** — t-test
- two samples of sizes n_1 and n_2

- two tail $H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2$
- right tail $H_0: \mu_1 < \mu_2, H_1: \mu_1 > \mu_2$
- left tail $H_0: \mu_1 > \mu_2, H_1: \mu_1 < \mu_2$



- **N normal dist.** — **AN**alysis **O**f **VA**riance
- N samples of sizes n_i

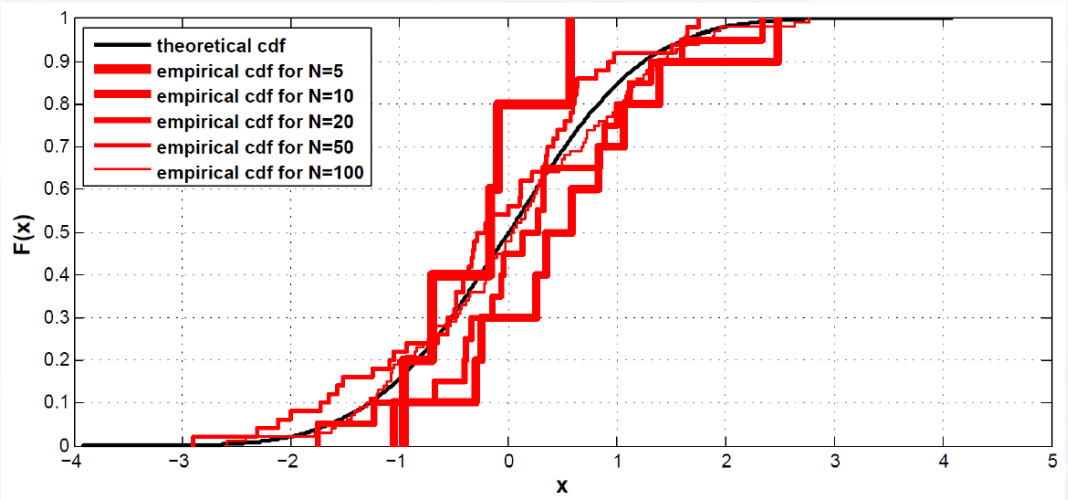
- $H_0: \mu_i = \mu_j \; \forall \; i, j$, $H_1: \mu_i \neq \mu_j$ for at least one pair i, j



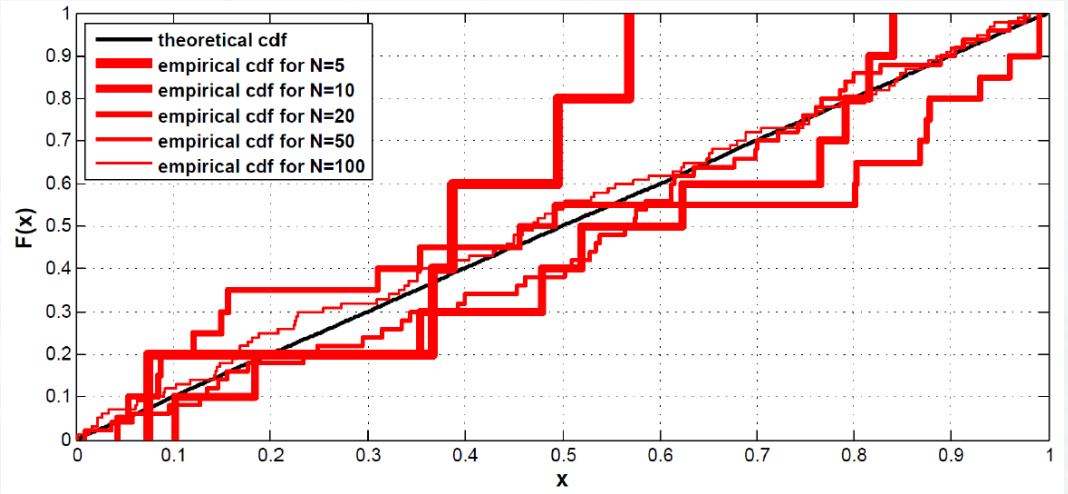
testing if a data set follows a particular distribution

Kolmogorov – **S**mirnov – test (KS test)

example: normal distribution:



example: uniform distribution:



Bayesian Model Selection/Testing:

testing if a data set follows a particular distribution

ranking tests (Wilcoxon)

…and many more…

They all generate a test statistic:

Z-value, t-value, F-value etc
→ from that: calculating a p-value

Berkeley UNIVERSITY OF CALIFORNIA

Bayesian Mod

the p-value

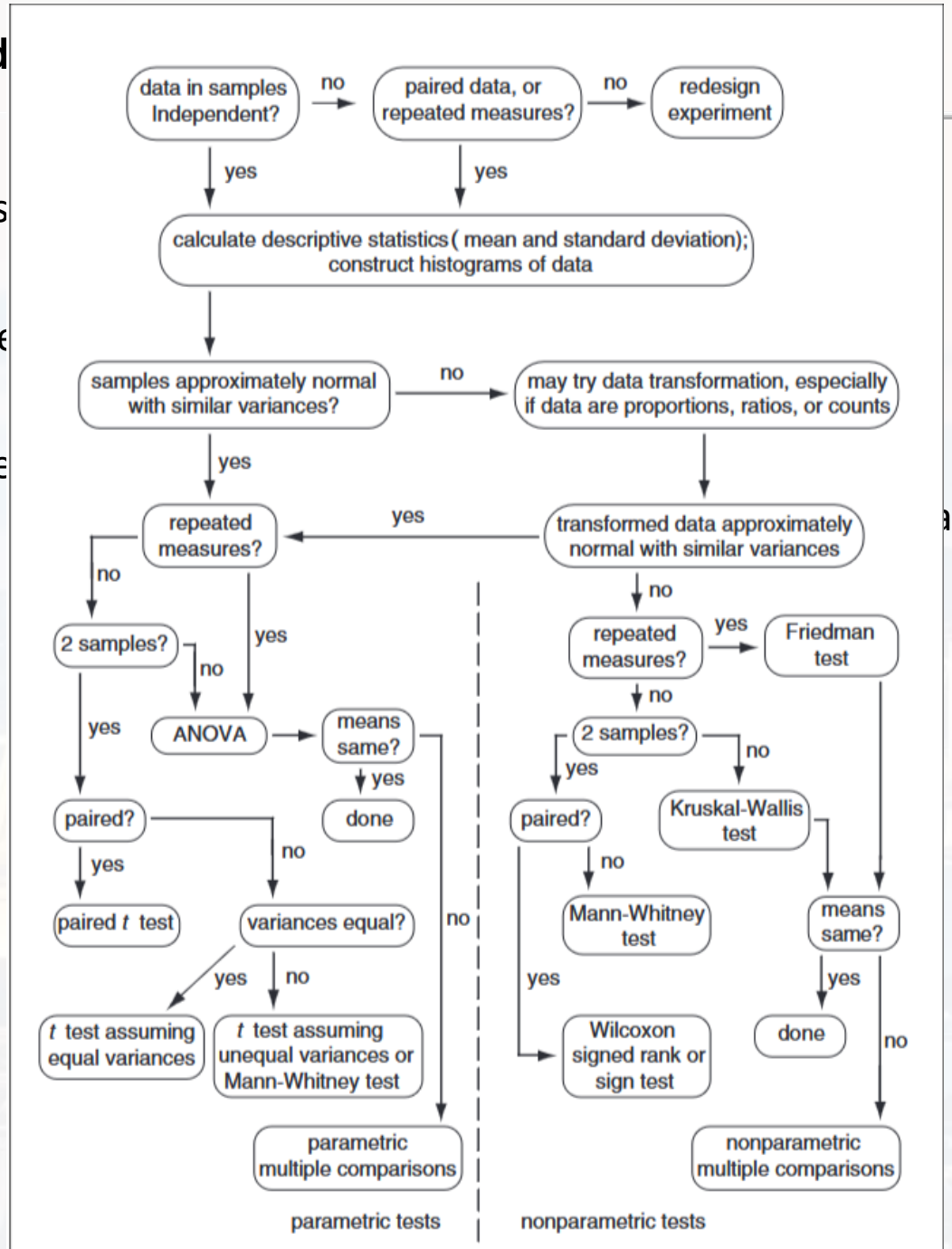
testing if a data s

ranking tests (Wilcoxon)

…and many more

They all generate

alue



Current Protocols in Molecular Biology (1998) A.3I.1-A.3I.22

Bayesian Model Selection/Testing

Outline

Standard Tests

- the p-value

Bayesian Model Testing

- idea
- curve fitting revisited
- comparing distributions

Bayesian Model Selection/Testing:

often, we have many competing models $\mathrm{M_i}$

goal: $\rho = \dfrac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)}$ odds ratio

$$= \frac{P(D|M_A,I)}{P(D|I)}\frac{P(M_A|I)}{} \cdot \frac{P(D|I)}{P(D|M_B,I)P(M_B|I)}$$

Bayes' theorem

idea
curve fitting revisited
comparing distributions

***D*:**	**data set**
***I*:**	**information**
M_A**:**	**model A**
M_B**:**	**model B**

Bayesian Model Selection/Testing:

often, we have many competing models $\mathrm{M_i}$

goal: $\rho = \dfrac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)}$ odds ratio

$$= \frac{P(D|M_A,I)}{P(D|I)}\frac{P(M_A|I)}{} \cdot \frac{P(D|I)}{P(D|M_B,I)P(M_B|I)}$$

idea
curve fitting revisited
comparing distributions

D:	data set
I:	information
M_A:	model A
M_B:	model B
$\{\alpha\}_i$:	all parameter of model M_i

marginalization:

$$P(D|M_i,I) = \int P(D,\{\alpha\}_i|M_i,I)\ d\Omega_{\{\alpha\}_i} \qquad P(x,y) = P(x|y)P(y)$$

$$= \int P(D|\{\alpha\}_i,M_i,I)\ P(\{\alpha\}_i|M_i,I)\ d\Omega_{\{\alpha\}_i}$$

$$= \int P(D|\{\alpha\}_i,M_i,I) \prod_j P(\alpha_{ij}|M_i,I)\ d\alpha_{ij}$$

assuming all α_{ij} are mutually independent **(Naïve Bayes)**

Bayesian Model Selection/Testing:

$$\boldsymbol{P(D|M_i,I)} = \int \underbrace{P(D|\{\alpha\}_i, M_i, I)}_{\text{likelihood function}} \prod_j \underbrace{P(\alpha_{ij}|M_i, I)}_{\text{prior of } \alpha_{ij}} \, d\alpha_{ij}$$

likelihood function
→ the actual model

prior of $\boldsymbol{\alpha_{ij}}$ BEVORE(!) measurement

maximum entropy without prior knowledge: $\frac{1}{\boldsymbol{\alpha_{ij}}(max) - \boldsymbol{\alpha_{ij}}(end)}$

idea
curve fitting revisited
comparing distributions

$\boldsymbol{D}$:	**data set**
$\boldsymbol{I}$:	**information**
$\boldsymbol{M_A}$:	**model A**
$\boldsymbol{M_B}$:	**model B**
$\{\boldsymbol{\alpha}\}_i$:	**all parameter of model $\boldsymbol{M_i}$**

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \underbrace{\frac{P(M_A)}{P(M_B)}}_{\text{prior probability of each model: maximum entropy → 1:1}} \cdot \underbrace{\frac{\int P(D|\{\alpha\}_A, M_A, I)\, d\alpha_{Aj}}{\int P(D|\{\alpha\}_B, M_B, I)\, d\alpha_{Bj}}}_{\text{likelihood function → the actual model}} \cdot \underbrace{\frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}}_{\text{Occam's Razor: simple models are preferred}}$$

Bayesian Model Selection/Testing:

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int P(D|\{\alpha\}_A, M_A, I)\, d\alpha_{Aj}}{\int P(D|\{\alpha\}_B, M_B, I)\, d\alpha_{Bj}} \cdot \frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}$$

idea
curve fitting revisited
comparing distributions

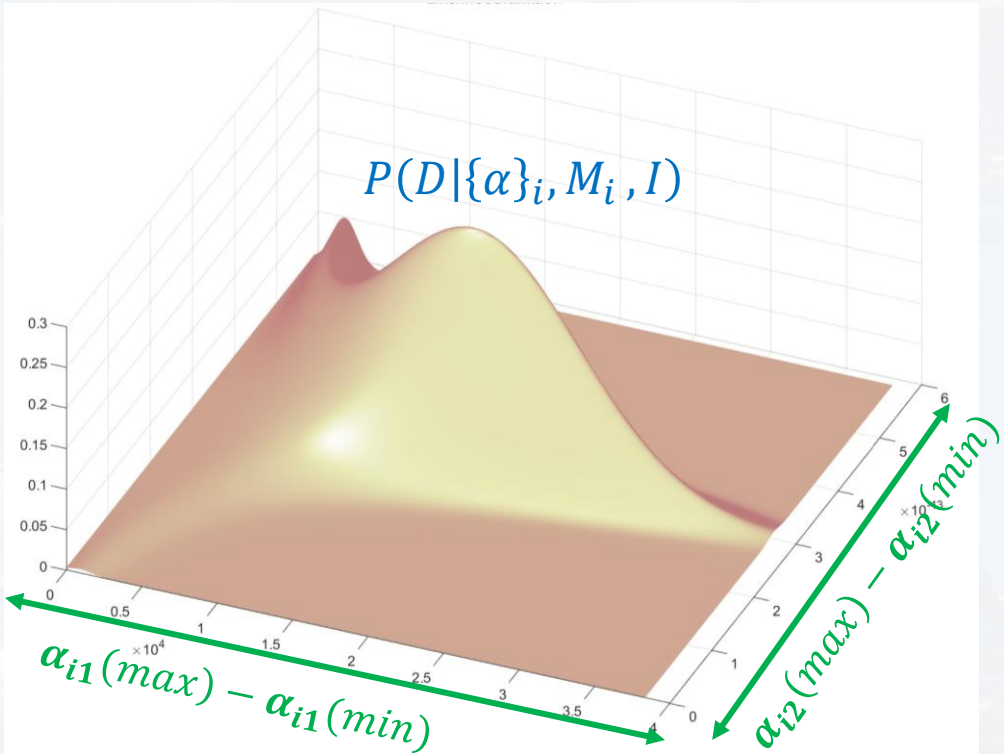
likelihood function
→ the actual model

Occam's Razor:
simple models are preferred

prior probability of each model: maximum entropy → 1:1

setting $\boldsymbol{\alpha_{ji}}(max) - \boldsymbol{\alpha_{ji}}(min)$ such that

- it covers the **entire integral** over $P(D|\{\alpha\}_i, M_i, I)$
- it makes **physical sense**
  (e. g. $\omega_{max} \approx$ Nyquist frequency,
  $\omega_{min} \approx 2/T_{obs}$)



Bayesian Model Selection/Testing:

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int P(D|\{\alpha\}_A, M_A, I)\, d\alpha_{Aj}}{\int P(D|\{\alpha\}_B, M_B, I)\, d\alpha_{Bj}} \cdot \frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}$$

idea
curve fitting revisited
comparing distributions

likelihood function
→ the actual model

Occam's Razor:
simple models are preferred

prior probability of each model: maximum entropy → 1:1

$$P(\alpha|D, M_i) = \frac{P(D|\alpha, M_i)P(\alpha|M_i)}{P(D|M_i)}$$

BPE: $P(\alpha|D) = \frac{P(D|\alpha)P(\alpha)}{P(D)}$

$P(D|M_i)$ is called **evidence** for the model M_i
(usually omitted for BPE)

Bayesian Model Selection/Testing

Outline

Standard Tests

- the p-value

Bayesian Model Testing

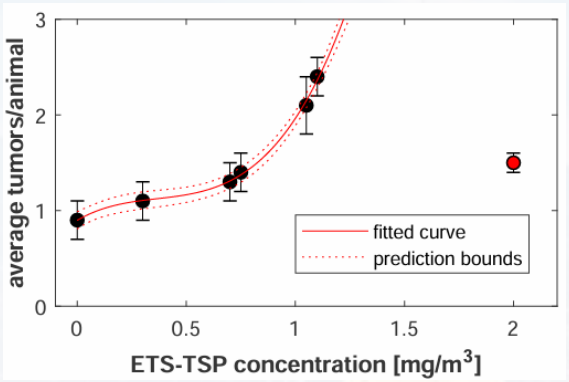
- idea
- curve fitting revisited
- comparing distributions

Bayesian Model Selection/Testing:

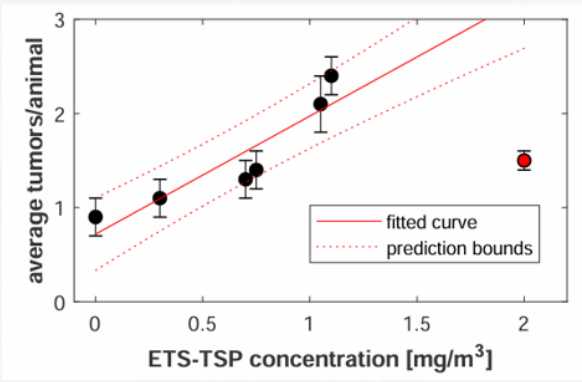
$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int P(D|\{\alpha\}_A, M_A, I)\, d\alpha_{Aj}}{\int P(D|\{\alpha\}_B, M_B, I)\, d\alpha_{Bj}} \cdot \frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}$$

idea
curve fitting revisited
comparing distributions

model $\mathrm{M_A}$



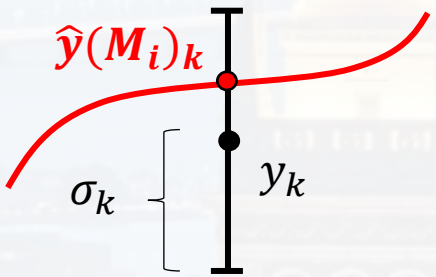
model $\mathrm{M_B}$



assumptions:

1) considerably good fit for each model M_i (otherwise, we wouldn't need Bayesian methods)

2) as before: y_k are drawn from a Gaussian $\mathcal{N}(y_k, \sigma_k)$



$$P(y_k|\{\alpha\}_i, M_i, I) \approx \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{1}{2}\frac{(y_k - \hat{y}(\boldsymbol{M_i})_k)^2}{\sigma_k^2}}$$

for $\sigma_k \ll |y_k|$

$\boldsymbol{y_k}$: **measured value**

$\boldsymbol{\sigma_k}$: **error**

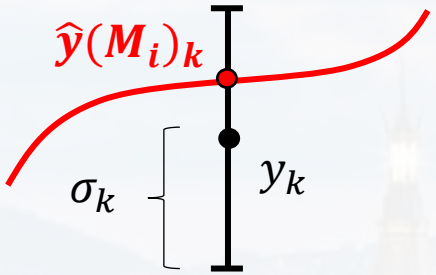
$\hat{\mathbf{y}}(\mathbf{M_i})_\mathbf{k}$: **model value (after fit)**

Bayesian Model Selection/Testing:

curve fitting revisited

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int P(D|\{\alpha\}_A, M_A, I)\, d\alpha_{Aj}}{\int P(D|\{\alpha\}_B, M_B, I)\, d\alpha_{Bj}} \cdot \frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}$$

idea
curve fitting revisited
comparing distributions



$$P(y_k|\{\alpha\}_i, M_i, I) \approx \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{1}{2}\frac{(y_k - \hat{y}(\boldsymbol{M_i})_k)^2}{\sigma_k^2}}$$

for $\sigma_k \ll |y_k|$

y_k: measured value
σ_k: error
$\hat{y}(\mathrm{M_i})_\mathrm{k}$: model value (after fit)

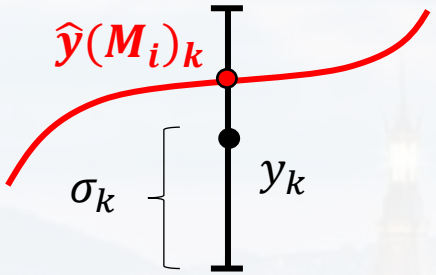
$$P(D|\{\alpha\}_i, M_i, I) = \prod_k P(y_k|\{\alpha\}_i, M_i, I) = \prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{1}{2}\frac{(y_k - \hat{y}(\boldsymbol{M_i})_k)^2}{\sigma_k^2}}$$

$$= \left(\prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\right) \cdot e^{-\frac{1}{2}\sum_k \frac{(y_k - \hat{y}(\boldsymbol{M_i})_k)^2}{\boldsymbol{\sigma_k^2}}}$$

$$= \left(\prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\right) \cdot e^{-\frac{1}{2}\boldsymbol{\chi_i^2}}$$

Bayesian Model Selection/Testing:

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int P(D|\{\alpha\}_A, M_A, I)\, d\alpha_{Aj}}{\int P(D|\{\alpha\}_B, M_B, I)\, d\alpha_{Bj}} \cdot \frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}$$



$$P(y_k|\{\alpha\}_i, M_i, I) \approx \frac{1}{\sqrt{2\pi\sigma_k^2}}\, e^{-\frac{1}{2}\frac{(y_k - \hat{y}(\boldsymbol{M_i})_k)^2}{\sigma_k^2}}$$

for $\sigma_k \ll |y_k|$

y_k: measured value
σ_k: error
$\hat{y}(\mathrm{M_i})_\mathrm{k}$: model value (after fit)

$$P(D|\{\alpha\}_i, M_i, I) = \left(\prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}}\right) \cdot e^{-\frac{1}{2}\chi_i^2}$$

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int e^{-\frac{1}{2}\chi_A^2}\, d\Omega_{\{\alpha\}_A}}{\int e^{-\frac{1}{2}\chi_B^2}\, d\Omega_{\{\alpha\}_B}} \cdot \frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}$$

Not only χ_i^2, but also model complexity (Occam's razor) and priors are important!

Bayesian Model Selection/Testing:

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int e^{-\frac{1}{2}\chi_A^2}\, d\Omega_{\{\alpha\}_A}}{\int e^{-\frac{1}{2}\chi_B^2}\, d\Omega_{\{\alpha\}_B}} \cdot \frac{\prod_j \boldsymbol{\alpha_{jB}}(max) - \boldsymbol{\alpha_{jB}}(min)}{\prod_j \boldsymbol{\alpha_{jA}}(max) - \boldsymbol{\alpha_{jA}}(min)}$$

note:

- like for MLE we found $\boldsymbol{\chi_i^2}$, if y_k are drawn from $\mathcal{N}(y_k, \sigma_k)$ but this time weighted with priors

- $\int e^{-\frac{1}{2}\chi_i^2}\, d\Omega_{\{\alpha\}_i}$ is usually hard to evaluate

  → numerically, or Monte-Carlo (see later), or

- $\chi_i^2 = \chi_i^2(\{\alpha\}_i)$ reaches a minimum for a certain set of $\{\alpha\}_i, \{\bar{\alpha}\}_i$

- 2nd order Taylor approx.:

$$\chi_i^2(\{\alpha\}_i) \approx \chi_i^2(\{\bar{\alpha}\}_i) + 0 + \frac{1}{2}(V_i - \bar{V}_i)^T\, \nabla\nabla\chi_i^2(\{\bar{\alpha}\}_i)(V_i - \bar{V}_i)$$

where V_i and $\bar{V}_i$ are vectors containing $\{\alpha\}_i$ and $\{\bar{\alpha}\}_i$, respectively

Bayesian Model Selection/Testing

Outline

Standard Tests

- the p-value

Bayesian Model Testing

- idea
- curve fitting revisited
- comparing distributions

Bayesian Model Selection/Testing:

problem: - often we want to know if two (or more) data sets D_m and D_n have been drawn from the same distribution (comparing μ or σ^2, like t-test or ANOVA)

- p – value **does not** answer the actual question!

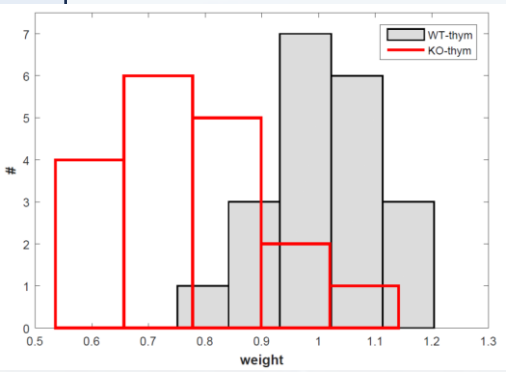
goal: - we want to be able to make the following statements:

- what is the probability that $\mu_m > \mu_n$
  $\sigma_m > \sigma_n$ … etc

- what is more likely: that D_m and D_n have been drawn from the same distribution or from two different distributions

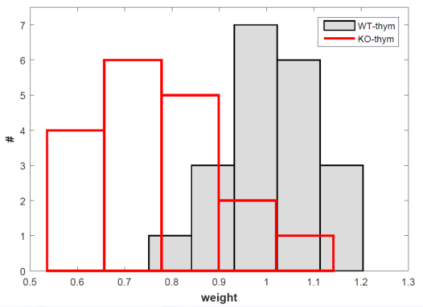
idea: calculating the **evidence** via **marginalization!**

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)}$$



example: $\mathcal{N}(y_k, \sigma_k)$

a) model A: data sets D_m and D_n have been drawn from the same distribution



$$P(D_m, D_n|\, A, I) = \iint P(D_m, D_n|\, \mu, \sigma, A, I)\, P(\mu, \sigma|A, I) d\mu d\sigma$$

if **max ent**:

$$P(\mu, \sigma|A, I) = \frac{1}{\mu_{max} - \mu_{min}} \frac{1}{\sigma_{max}}$$

likelihood function
→ the actual model

D_m: N_m data points
D_n: N_n data points

$$P(D_m, D_n|\, \mu, \sigma, A, I) = \prod_{i=1}^{N_{tot}} P_i(D_m, D_n|\, \mu, \sigma, A, I)$$

$$N_{tot} = N_n + N_m$$

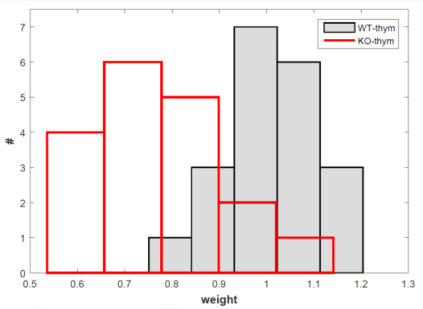
$$= \frac{1}{(2\pi\sigma^2)^{N_{tot}/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N_{tot}}(x_i-\mu)^2}$$

Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

a) model A: data sets D_m and D_n have been drawn from the same distribution

$$P(D_m, D_n | A, I) = \frac{1}{\mu_{max} - \mu_{min}} \frac{1}{\sigma_{max}} \iint \frac{1}{(2\pi\sigma^2)^{N_{tot}/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N_{tot}}(x_i-\mu)^2} \, d\mu d\sigma$$



again, the integral is usually hard to evaluate:

→ numerically, or

→ Taylor approximation for large N_{tot} around $\hat{\sigma}^2$ and $\hat{\mu}$

D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

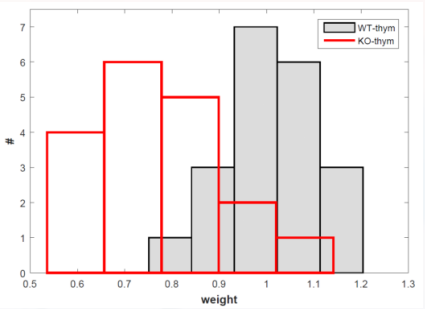
where

$$\hat{\sigma}^2 = \frac{1}{N_{tot}} \sum_{i=1}^{N_{tot}} (x_i - \hat{\mu})^2 \quad \text{and} \quad \hat{\mu} = \frac{1}{N_{tot}} \sum_{i=1}^{N_{tot}} {x_i}^2$$

example: $\mathcal{N}(y_k, \sigma_k)$

a) model A: data sets D_m and D_n have been drawn from the same distribution

$$P(D_m, D_n | A, I) = \frac{1}{\mu_{max} - \mu_{min}} \frac{1}{\sigma_{max}} \iint \frac{1}{(2\pi\sigma^2)^{N_{tot}/2}} \cdot e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{N_{tot}}(x_i-\mu)^2} \, d\mu d\sigma$$



again, the integral is usually hard to evaluate:

→ numerically, or

→ Taylor approximation for large $\boldsymbol{N_{tot}}$ around $\hat{\sigma}^2$ and $\hat{\mu}$

$\boldsymbol{D_m}$: $\boldsymbol{N_m}$ **data points**
$\boldsymbol{D_n}$: $\boldsymbol{N_n}$ **data points**
$\boldsymbol{N_{tot} = N_n + N_m}$

$$\hat{\mu} = \frac{1}{N_{tot}} \sum_{i=1}^{N_{tot}} x_i^2$$

$$P(D_m, D_n | A, I) \approx \frac{1}{\sqrt{2}} \frac{\left(\hat{\sigma}\sqrt{2\pi}\right)^{2-N_{tot}} e^{-N_{tot}/2}}{(\mu_{max} - \mu_{min})\, \sigma_{max}\, N_{tot}}$$

$$\hat{\sigma}^2 = \frac{1}{N_{tot}} \sum_{i=1}^{N_{tot}} (x_i - \hat{\mu})^2$$

Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

b) model B: data sets D_m and D_n have been drawn from **different** distributions

$$P(D_m, D_n | B, I) = P(D_m | B, I)\, P(D_n | B, I)$$

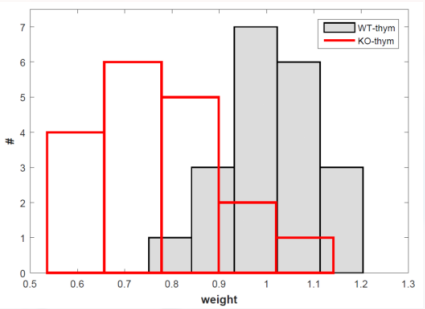
in general:

$$P(\{D_k\} | B, I) = \prod_{k=1}^{K} P(D_k | B, I)$$

as before: marginalization over evidence term **(assuming same ranges for μ and σ)**

$$P(D_k | B, I) = \frac{1}{\mu_{max} - \mu_{min}} \frac{1}{\sigma_{max}} \iint \frac{1}{[2\pi\, \sigma^2(k)]^{N_k/2}} \cdot e^{-\frac{1}{2\sigma^2(k)} \sum_{i=1}^{N_k} (x_i - \mu(k))^2}\, d\mu(k)\, d\sigma(k)$$

now: comparing odds ratios!



D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

now: comparing odds ratios!

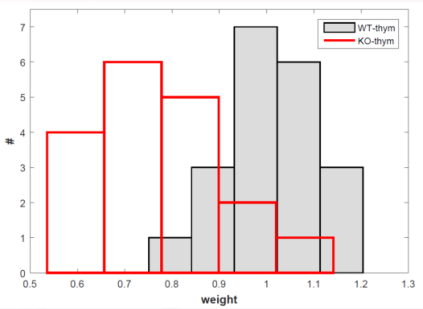
$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(D_m, D_n|\, A, I)}{P(D_m, D_n|\, B, I)} \frac{P(A|I)}{P(B|I)}$$

$$= \frac{(\mu_{max} - \mu_{min})\, \sigma_{max}}{\sqrt{2\pi}} \cdot \frac{N_m\, N_n}{N_{tot}} \cdot \frac{(\hat{\sigma})^{2-\boldsymbol{N_{tot}}}}{(\hat{\sigma}_n)^{2-\boldsymbol{N_n}}\, (\hat{\sigma}_m)^{2-\boldsymbol{N_m}}} \cdot \frac{P(A|I)}{P(B|I)}$$

for *K* different distributions:

$$P(\{D_k\}|\, B, I) = \prod_{k=1}^{K} P(D_k|\, B, I)$$

$$\rho = \left[\frac{(\mu_{max} - \mu_{min})\, \sigma_{max}}{\sqrt{2\pi}}\right]^{K-1} \cdot \frac{(\hat{\sigma})^{2-N_{tot}}}{N_{tot}} \cdot \prod_{k=1}^{K} \frac{N_k}{(\hat{\sigma}_k)^{2-N_k}} \cdot \frac{P(A|I)}{P(B|I)}$$



$\boldsymbol{D_m}$: $\boldsymbol{N_m}$ data points
$\boldsymbol{D_n}$: $\boldsymbol{N_n}$ data points

$\boldsymbol{N_{tot} = N_n + N_m}$

Bayesian Model Selection/Testing:

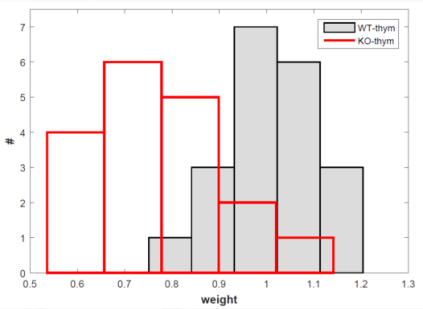
example: $\mathcal{N}(y_k, \sigma_k)$

now: comparing odds ratios!

$$\rho = \frac{P(\mathrm{M_A}|D,I)}{P(\mathrm{M_B}|D,I)} = \frac{P(D_m, D_n|\, A, I)}{P(D_m, D_n|\, B, I)} \frac{P(A|I)}{P(B|I)}$$

$$= \frac{(\mu_{max} - \mu_{min})\, \sigma_{max}}{\sqrt{2\pi}} \cdot \frac{N_m\, N_n}{N_{tot}} \cdot \frac{(\hat{\sigma})^{2-N_{tot}}}{(\hat{\sigma}_n)^{2-N_n}\, (\hat{\sigma}_m)^{2-N_m}} \cdot \frac{P(A|I)}{P(B|I)}$$

$$\rho = \left[\frac{(\mu_{max}-\mu_{min})\, \sigma_{max}}{\sqrt{2\pi}}\right]^{K-1} \cdot \frac{(\hat{\sigma})^{2-N_{tot}}}{N_{tot}} \cdot \prod_{k=1}^{K} \frac{N_k}{(\hat{\sigma}_k)^{2-N_k}} \cdot \frac{P(A|I)}{P(B|I)}$$



D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

note: - can be extended to any alternative distributions for $\mathrm{M_B}$

- alternative to a **two-tailed t-test** (if model is $\mathcal{N}(y_k, \sigma_k)$)

example: $\mathcal{N}(y_k, \sigma_k)$

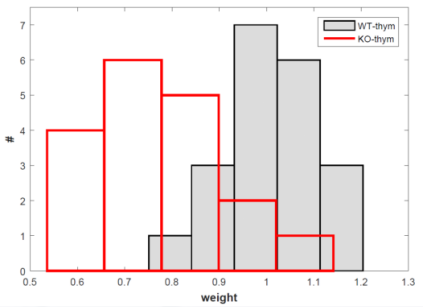
now: alternative to a **one-tailed t-test** (if model is $\mathcal{N}(y_k, \sigma_k)$)

Given D_1 and D_2, what is the probability that $\mu_1 > \mu_2$?

μ_1 can have any value, but once set, it is a constrain for μ_2

$$P(\mu_1 > \mu_2|D_1, D_2, I) = \int_{0\ or -\infty}^{\infty} \int_{0}^{\mu_1} P(\mu_1, \mu_2|D_1, D_2, I)\, d\mu_2\, d\mu_1$$

$$= \int_{0\ or -\infty}^{\infty} \int_{0}^{\mu_1} P(\mu_1|D_1, I)\, P(\mu_2|D_2, I) d\mu_2\, d\mu_1$$



D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

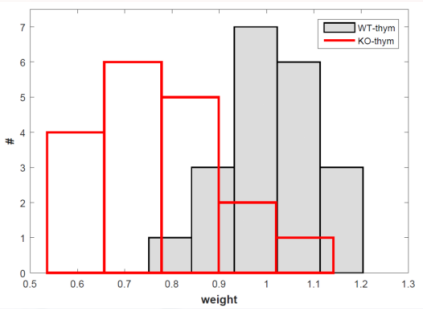
Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

Given D_1 and D_2, what is the probability that $\mu_1 > \mu_2$?

$$P(\mu_1 > \mu_2 | D_1, D_2, I) = \int_{0\ or -\infty}^{\infty} \int_{0}^{\mu_1} P(\mu_1, \mu_2 | D_1, D_2, I)\, d\mu_1\, d\mu_2$$

$$= \int_{0\ or -\infty}^{\infty} \int_{0}^{\mu_1} P(\mu_1 | D_1, I)\, P(\mu_2 | D_2, I) d\mu_1\, d\mu_2$$



D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

we only compare means: marginalization over σ_k

$$P(\mu_k | D_k, I) = \int_0^{\infty} P(\mu_k, \sigma_k | D_k, I)\, d\sigma_k$$

$$\sim \int_0^{\infty} P(D_k | \mu_k, \sigma_k, I) P(\mu_k, \sigma_k | I)\, d\sigma_k$$

max ent:

$$P(\mu_k, \sigma_k | D_k, I) = \frac{P(D_k | \mu_k, \sigma_k, I)\, P(\mu_k, \sigma_k | I)}{P(D_k, I)}$$

Bayes' theorem

Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

Given D_1 and D_2, what is the probability that $\mu_1 > \mu_2$?

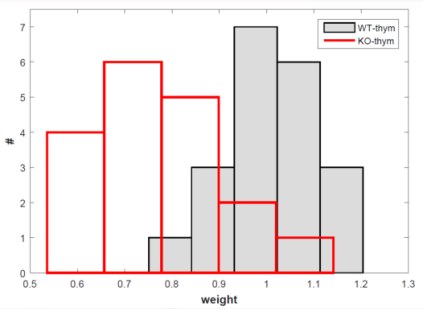
$$P(\mu_1 > \mu_2 | D_1, D_2, I) = \int_{0\ or -\infty}^{\infty} \int_{0}^{\mu_1} P(\mu_1, \mu_2 | D_1, D_2, I)\, d\mu_1\, d\mu_2$$

$$= \int_{0\ or -\infty}^{\infty} \int_{0}^{\mu_1} P(\mu_1 | D_1, I)\, P(\mu_2 | D_2, I) d\mu_1\, d\mu_2$$

$$P(\mu_k | D_k, I) \sim \int_{0}^{\infty} P(D_k |\, \mu_k, \sigma_k, I) P(\mu_k, \sigma_k | I)\, d\sigma_k$$

likelihood function
→ the actual model

$$P(D_k |\, \mu_k, \sigma_k, I) = \frac{1}{(2\pi {\sigma_k}^2)^{N_k/2}} \cdot e^{-\frac{1}{2{\sigma_k}^2} \sum_{i=1}^{N_k} (x_i - \mu_k)^2}$$



D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

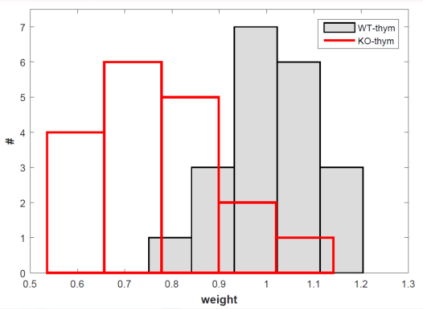
Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

Given D_1 and D_2, what is the probability that $\mu_1 > \mu_2$?

again, approximating the integral for large *N*

$$P(\mu_1 > \mu_2 | D_1, D_2, I) \approx \frac{1}{S_{tot}\sqrt{2\pi}} \int_0^\infty \exp\left[-\frac{1}{2\,S_{tot}^2}(\zeta - \hat{z})^2\right] d\zeta$$



D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

where:

$$S_{tot}^2 = \frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}$$

$$\hat{z} = \hat{\mu}_1 - \hat{\mu}_2$$

$$\zeta = \mu_1 - \mu_2$$

$$\hat{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_{i,k} \qquad S_k^2 = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}\left(x_{i,k} - \hat{\mu}_k\right)^2$$

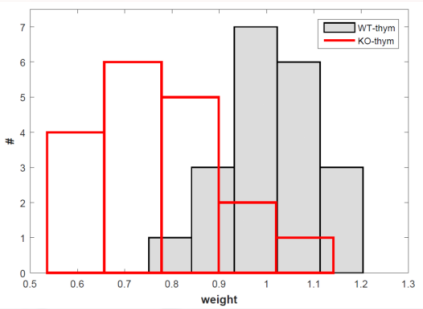
Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

Given D_1 and D_2, what is the probability that $\mu_1 > \mu_2$?

again, approximating the integral for large *N*

$$P(\mu_1 > \mu_2 | D_1, D_2, I) \approx \frac{1}{S_{tot}\sqrt{2\pi}} \int_0^\infty \exp\left[-\frac{1}{2\,S_{tot}^2}(\zeta - \hat{z})^2\right] d\zeta$$



where:

$$S_{tot}^2 = \frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}$$

$$\hat{\mu}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} x_{i,k}$$

$$\hat{z} = \hat{\mu}_1 - \hat{\mu}_2$$

$$\zeta = \mu_1 - \mu_2$$

$$S_k^2 = \frac{1}{N_k - 1}\sum_{i=1}^{N_k}\left(x_{i,k} - \hat{\mu}_k\right)^2$$

D_m: N_m data points
D_n: N_n data points

$N_{tot} = N_n + N_m$

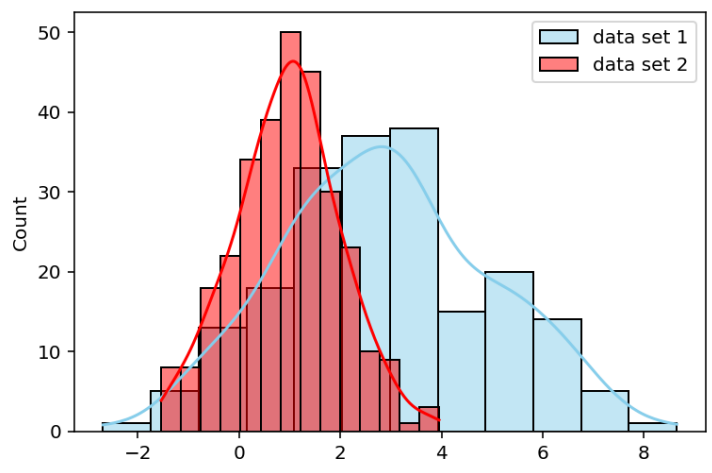
note: - more detailed derivation: Sivia, "Data Analysis", page 53

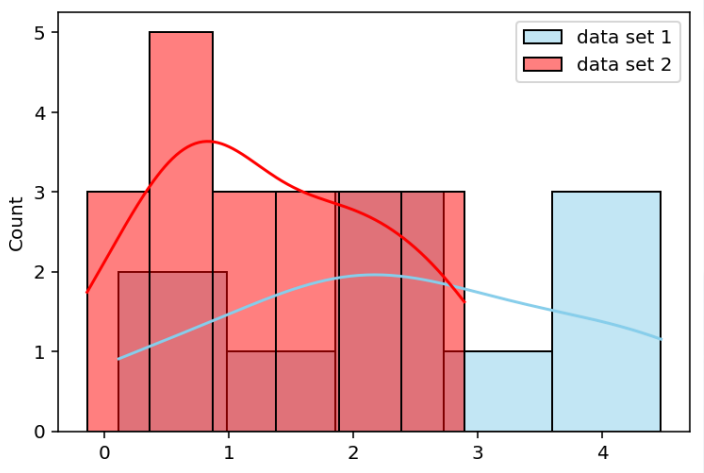
- changing μ_1 with μ_2 changes $P(\mu_1 > \mu_2 | D_1, D_2, I)$ to $1 - P(\mu_1 > \mu_2 | D_1, D_2, I)$

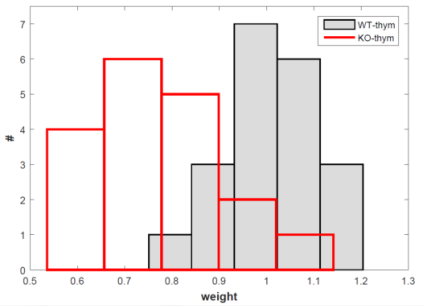
Bayesian Model Selection/Testing:

example: $\mathcal{N}(y_k, \sigma_k)$

Check out: `mean_var_analysis_example.py`







D_m: N_m data points
D_n: N_n data points
$N_{tot} = N_n + N_m$

```
M.ModelSelection()
```

```
M.T_Test()
(2.144, 0.060)
```

p – value

t – value: $t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}}$

```
M.ModelSelection()
0.711
```

$$\rho = \frac{P(\mathrm{M_A} = from\ one\ pdf | D, I)}{P(\mathrm{M_B} = from\ two\ pdfs | D, I)}$$

```
M.P_Means()
(0.99, 0.01)
```

$P(\mu_1 > \mu_2 | D_1, D_2, I)$ and $P(\mu_1 < \mu_2 | D_1, D_2, I)$

Bayesian Model Selection/Testing

Thank you very much for your attention!