

## Lecture 3:

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



Markus Hohle

University California, Berkeley

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
<b>Module 3</b>	<b>MLE, Lin Regression, Model selection: Comparing Distributions</b>
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)



## Outline

### Standard Tests

- the p-value

### Bayesian Model Testing

- idea
- curve fitting revisited
- comparing distributions





## Outline

### Standard Tests

- the p-value

### Bayesian Model Testing

- idea
- curve fitting revisited
- comparing distributions

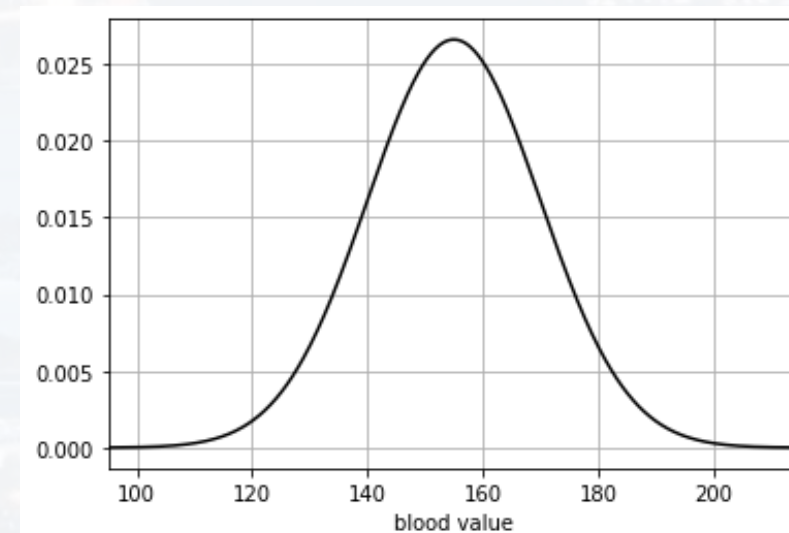


standard frequentist way (not Bayesian):

- 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  $\alpha$

**example:**

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





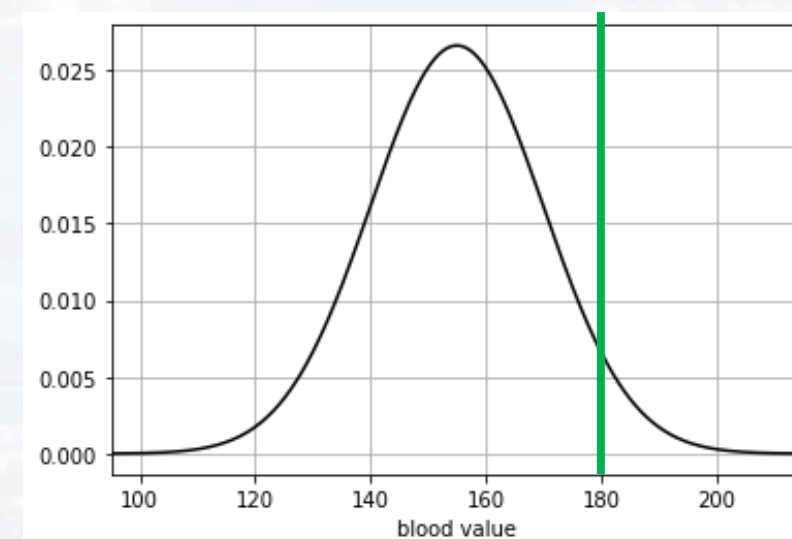
## standard frequentist way (not Bayesian):

- 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  $\alpha$

### 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  **$x_0 = 180$**







## 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  $\alpha$

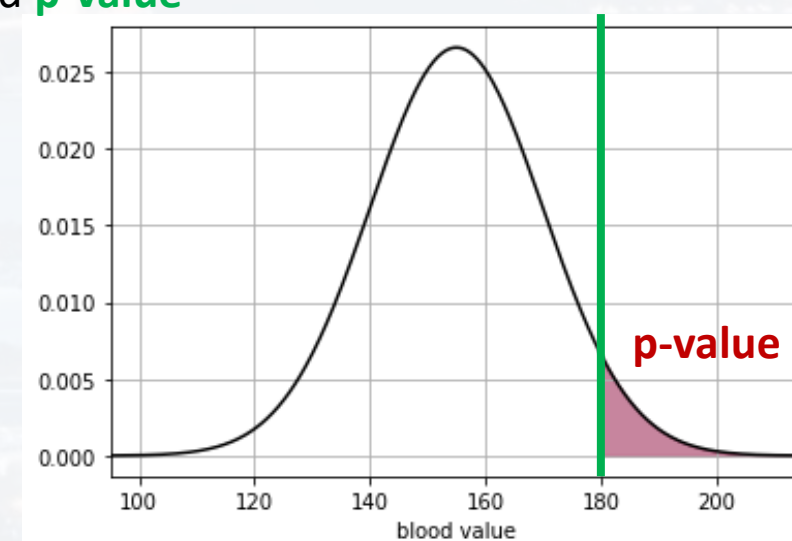
### example:

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  $\alpha$

**example:**

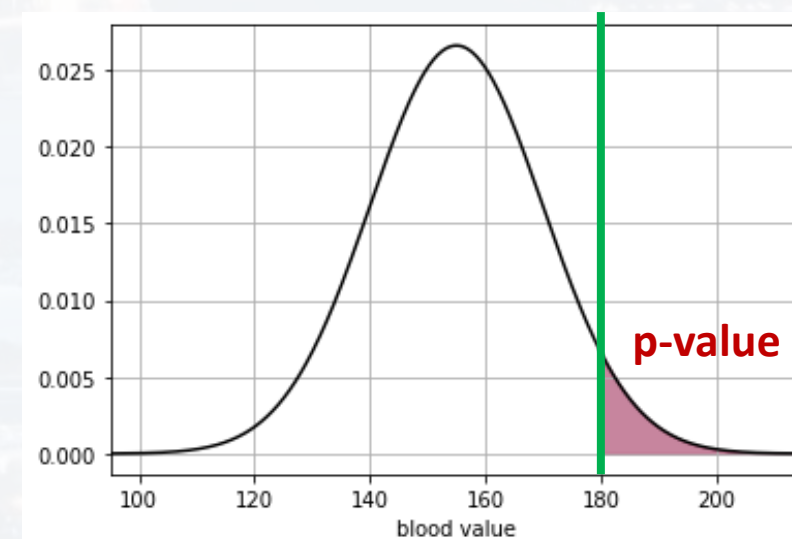
1) a healthy person has a blood value that follows a **normal distribution** with  $\mu = 155, \sigma = 15$ , i. e.  **$H_0: N(\mu = 155, \sigma = 15)$**

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

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

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

→ alternative hypothesis  **$H_1$ : not healthy**







## 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  **$x_0 = 180$**

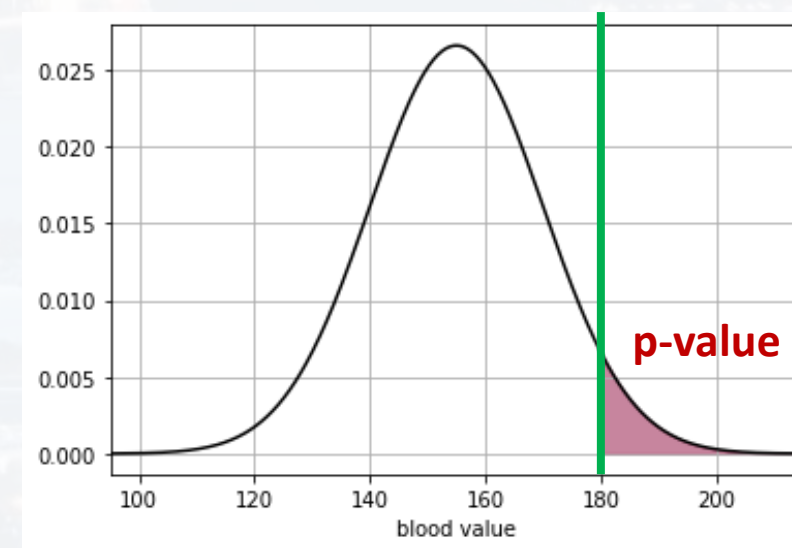
3) probability  **$P(x \geq 180|H_0) = 0.048$**  called **p-value**

4) for  **$\alpha = 0.05$** ,  $H_0$  is **rejected**, i. e. patient is

**not healthy**

→ alternative hypothesis  **$H_1$ : not healthy**

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





## 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  $\alpha$

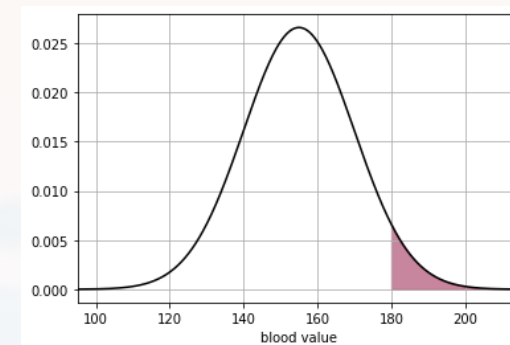
## caveats:

- 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  $\alpha$  for accepting/rejecting  $H_0$  is **arbitrary**
- we are aiming on disproving a hypothesis, by assuming it is true, **without** leading to a contradiction



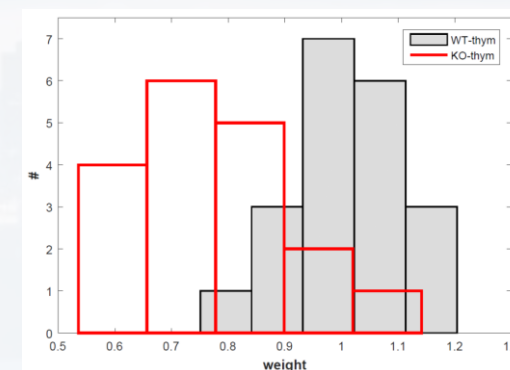
- data point  $x$  versus **normal**
- $H_0: x \in N(\mu, \sigma)$

Z-test



- **normal** versus another **normal**
- two samples of sizes  $n_1$  and  $n_2$

t-test

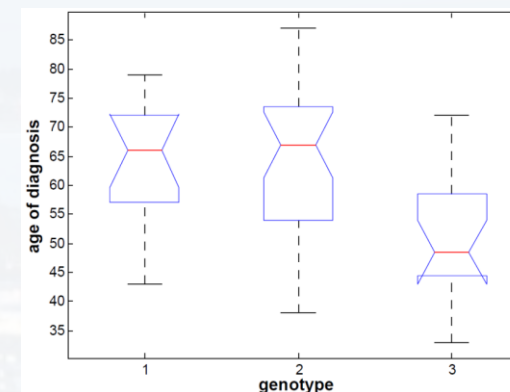


- two tail  $H_0: \mu_1 = \mu_2, H_1: \mu_1 \neq \mu_2$
- right tail  $H_0: \mu_1 < \mu_2, H_1: \mu_1 > \mu_2$
- left tail  $H_0: \mu_1 > \mu_2, H_1: \mu_1 < \mu_2$

- **N normal dist.**

ANalysis Of VAriance

- 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  $i, j$



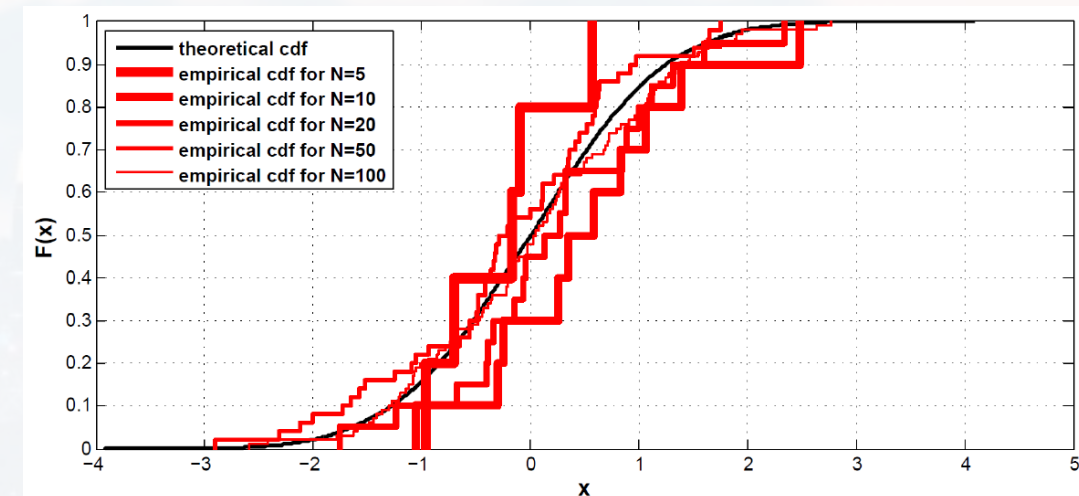




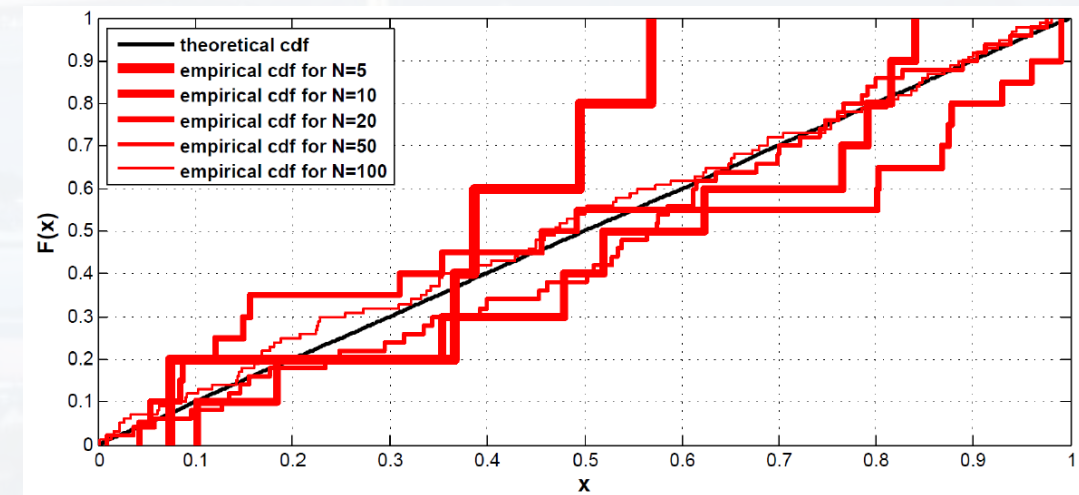
testing if a data set follows a particular distribution

Kolmogorov – Smirnov – test (KS test)

example: normal distribution:



example: uniform distribution:





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

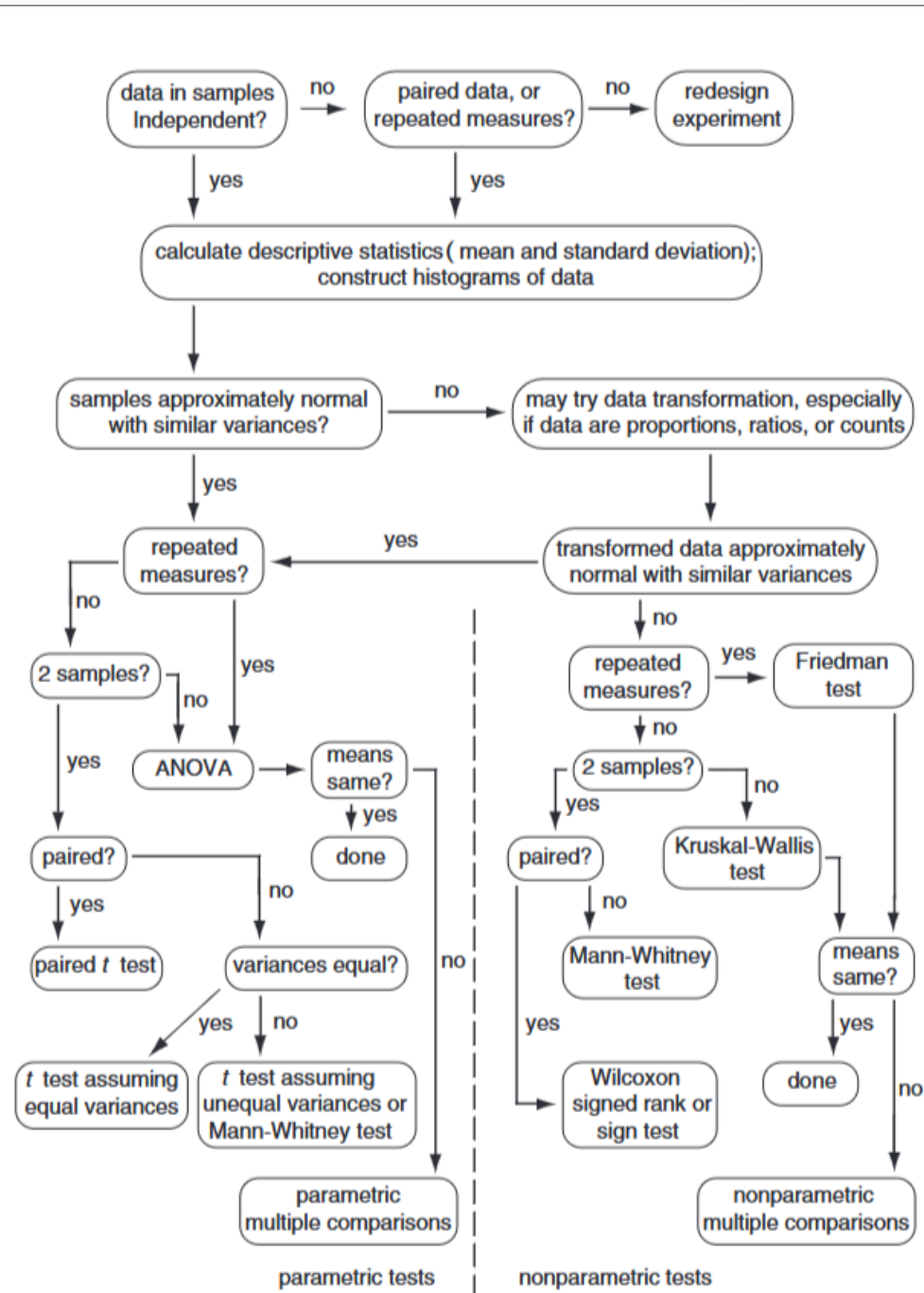
→ from that: calculating a p-value



testing if a data s

...and many more

They all generate



the p-value

ranking tests (Wilcoxon)

value





## Outline

### Standard Tests

- the p-value

### Bayesian Model Testing

- idea
- curve fitting revisited
- comparing distributions



often, we have many competing models  $M_i$

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

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

Bayes' theorem

idea

curve fitting revisited  
comparing distributions

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



often, we have many competing models  $M_i$

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

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

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:

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

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

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

assuming all  $\alpha_{ij}$  are mutually independent (**Naïve Bayes**)





$$P(D|M_i, I) = \int \underbrace{P(D|\{\alpha\}_i, M_i, I)}_{\text{likelihood function} \rightarrow \text{the actual model}} \prod_j \underbrace{P(\alpha_{ij}|M_i, I)}_{\text{prior of } \alpha_{ij} \text{ BEVORE(!) measurement}}$$

likelihood function  
→ the actual model

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

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

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 $M_i$

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

likelihood function  
→ the actual model

**Occam's Razor:**  
simple models are preferred

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



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

idea

curve fitting revisited  
comparing distributions

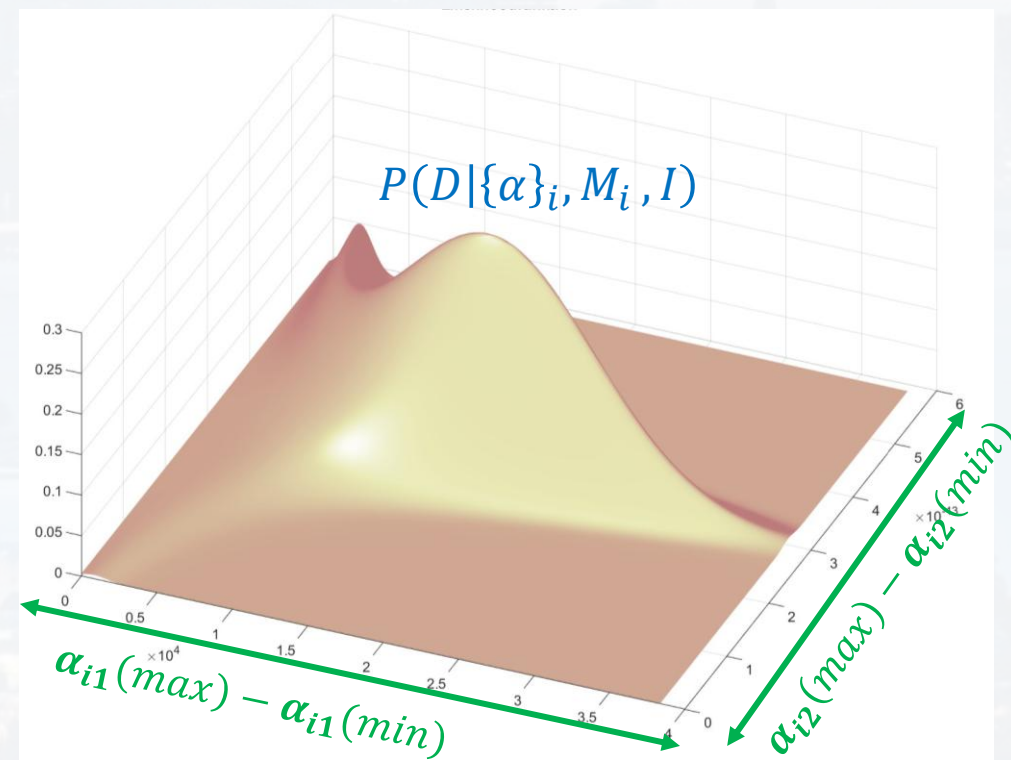
likelihood function  
→ the actual model

**Occam's Razor:**  
simple models are preferred

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

setting  $\alpha_{ji}(max) - \alpha_{ji}(min)$  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,  
 $\omega_{min} \approx 2/T_{obs}$ )





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

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

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

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

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





## Outline

### Standard Tests

- the p-value

### Bayesian Model Testing

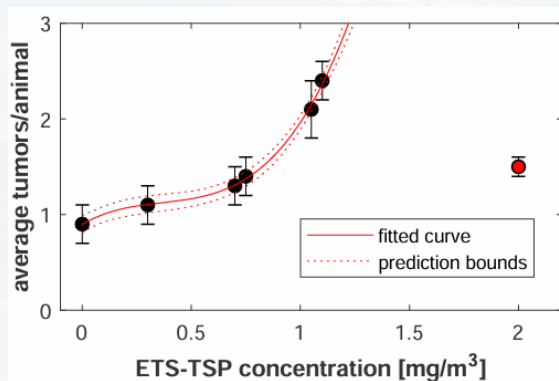
- idea
- curve fitting revisited
- comparing distributions



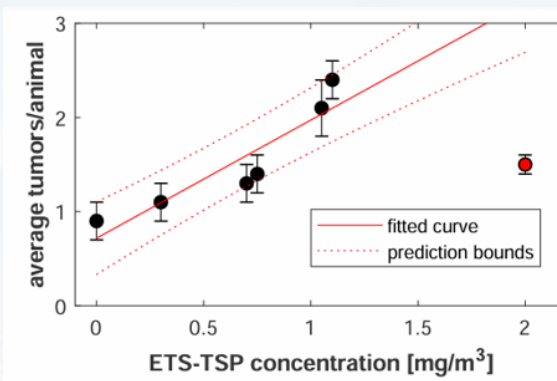
$$\rho = \frac{P(M_A|D, I)}{P(M_B|D, I)} = \frac{P(M_A)}{P(M_B)} \cdot \frac{\int P(D|\{\alpha\}_A, M_A, I) d\alpha_{Aj}}{\int P(D|\{\alpha\}_B, M_B, I) d\alpha_{Bj}} \cdot \frac{\prod_j \alpha_{jB}(\max) - \alpha_{jB}(\min)}{\prod_j \alpha_{jA}(\max) - \alpha_{jA}(\min)}$$

idea  
curve fitting revisited  
comparing distributions

model  $M_A$

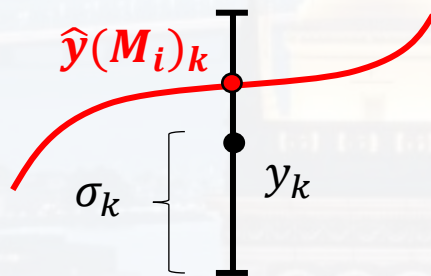


model  $M_B$



## 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  $\mathcal{N}(y_k, \sigma_k)$



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

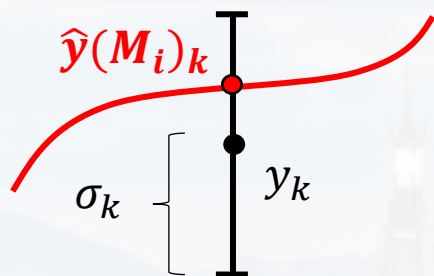
for  $\sigma_k \ll |y_k|$

$y_k$ : measured value  
 $\sigma_k$ : error  
 $\hat{y}(M_i)_k$ : model value (after fit)



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

idea  
curve fitting revisited  
comparing distributions



$y_k$ : measured value  
 $\sigma_k$ : error  
 $\hat{y}(M_i)_k$ : model value (after fit)

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

for  $\sigma_k \ll |y_k|$

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

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

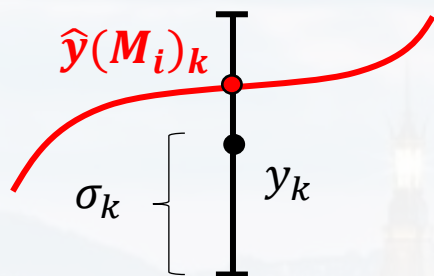
$$= \left( \prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \right) \cdot e^{-\frac{1}{2} \chi_i^2}$$





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

idea  
curve fitting revisited  
comparing distributions



$y_k$ : measured value  
 $\sigma_k$ : error  
 $\hat{y}(M_i)_k$ : model value (after fit)

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

for  $\sigma_k \ll |y_k|$

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

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

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



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

idea  
curve fitting revisited  
comparing distributions

**note:** - like for MLE we found  $\chi_i^2$ , 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  $\{\alpha\}_i, \{\bar{\alpha}\}_i$

- 2<sup>nd</sup> 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



## Outline

### Standard Tests

- the p-value

### Bayesian Model Testing

- idea
- curve fitting revisited
- comparing distributions





**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  $\mu$  or  $\sigma^2$ , like t-test or ANOVA)

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

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

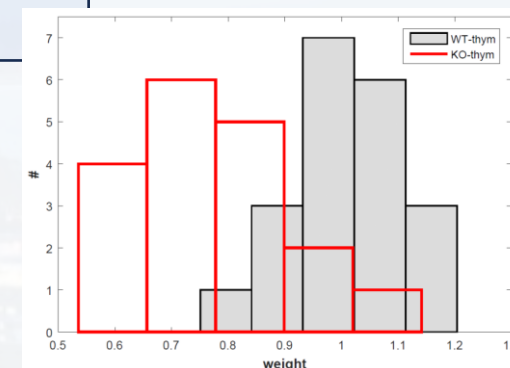
- what is the probability that  $\mu_m > \mu_n$   
 $\sigma_m > \sigma_n$  ... etc

- 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!**

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

idea  
curve fitting revisited  
comparing distributions





**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:**

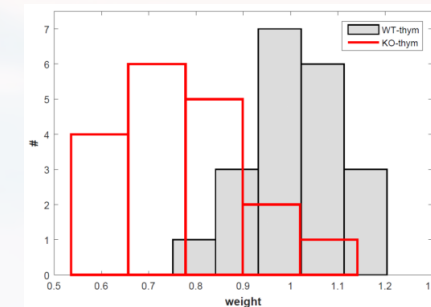
$$P(\mu, \sigma | A, I) = \frac{1}{\mu_{\max} - \mu_{\min}} \frac{1}{\sigma_{\max}}$$

likelihood function  
→ the actual model

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

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

idea  
curve fitting revisited  
comparing distributions



$D_m$ :  $N_m$  data points

$D_n$ :  $N_n$  data points

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



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

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

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

again, the integral is usually hard to evaluate:

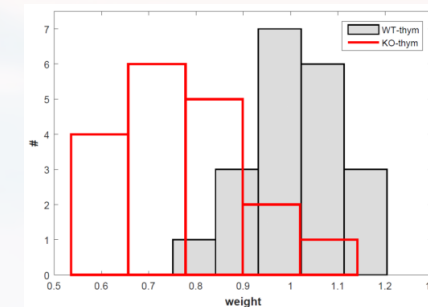
→ numerically, or

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

where

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

idea  
curve fitting revisited  
comparing distributions



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





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

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

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

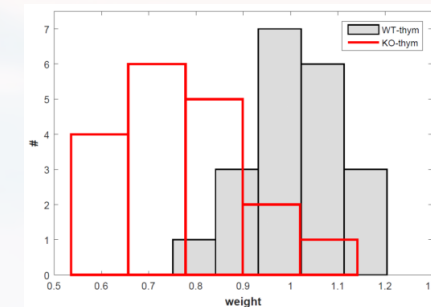
again, the integral is usually hard to evaluate:

→ numerically, or

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

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

idea  
curve fitting revisited  
comparing distributions



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

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

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



**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:

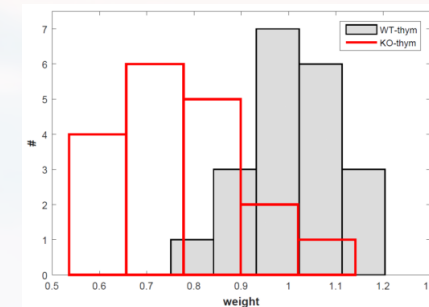
$$P(\{D_k\} | B, I) = \prod_{k=1}^K P(D_k | B, I)$$

as before: marginalization over evidence term (assuming same ranges for  $\mu$  and  $\sigma$ )

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

now: comparing odds ratios!

idea  
curve fitting revisited  
comparing distributions



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



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

now: comparing odds ratios!

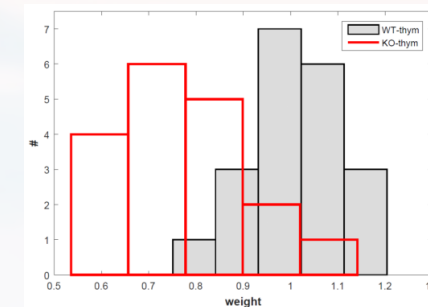
$$\begin{aligned} \rho &= \frac{P(M_A|D, I)}{P(M_B|D, I)} = \frac{P(D_m, D_n | A, I)}{P(D_m, D_n | B, I)} \frac{P(A|I)}{P(B|I)} \\ &= \frac{(\mu_{max} - \mu_{min}) \sigma_{max}}{\sqrt{2}\pi} \cdot \frac{N_m N_n}{N_{tot}} \cdot \frac{(\hat{\sigma})^{2-N_{tot}}}{(\hat{\sigma}_n)^{2-N_n} (\hat{\sigma}_m)^{2-N_m}} \cdot \frac{P(A|I)}{P(B|I)} \end{aligned}$$

for  $K$  different distributions:

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

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

idea  
curve fitting revisited  
comparing distributions



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



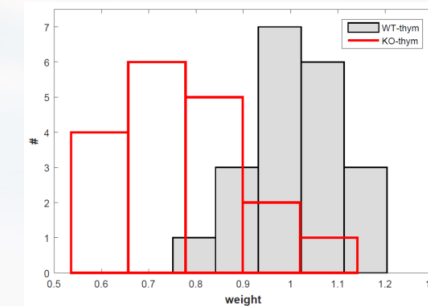


**example:**  $\mathcal{N}(y_k, \sigma_k)$

now: comparing odds ratios!

$$\begin{aligned} \rho &= \frac{P(M_A|D, I)}{P(M_B|D, I)} = \frac{P(D_m, D_n | A, I)}{P(D_m, D_n | B, I)} \frac{P(A|I)}{P(B|I)} \\ &= \frac{(\mu_{max} - \mu_{min}) \sigma_{max}}{\sqrt{2}\pi} \cdot \frac{N_m N_n}{N_{tot}} \cdot \frac{(\hat{\sigma})^{2-N_{tot}}}{(\hat{\sigma}_n)^{2-N_n} (\hat{\sigma}_m)^{2-N_m}} \cdot \frac{P(A|I)}{P(B|I)} \\ \rho &= \left[ \frac{(\mu_{max} - \mu_{min}) \sigma_{max}}{\sqrt{2}\pi} \right]^{K-1} \cdot \frac{(\hat{\sigma})^{2-N_{tot}}}{N_{tot}} \cdot \prod_{k=1}^K \frac{N_k}{(\hat{\sigma}_k)^{2-N_k}} \cdot \frac{P(A|I)}{P(B|I)} \end{aligned}$$

idea  
curve fitting revisited  
comparing distributions



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

**note:**

- can be extended to any alternative distributions for  $M_B$
- 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  $\mu_1 > \mu_2$ ?

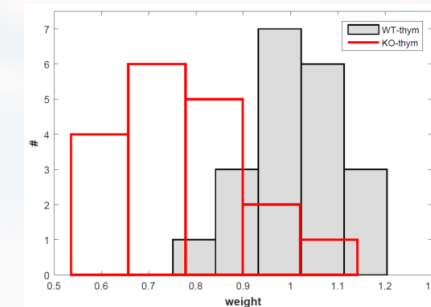


$\mu_1$  can have any value, but once set, it is a constrain for  $\mu_2$

$$P(\mu_1 > \mu_2 | D_1, D_2, I) = \int_{0 \text{ or } -\infty}^{\infty} \int_0^{\mu_1} P(\mu_1, \mu_2 | D_1, D_2, I) d\mu_2 d\mu_1$$

$$= \int_{0 \text{ or } -\infty}^{\infty} \int_0^{\mu_1} P(\mu_1 | D_1, I) P(\mu_2 | D_2, I) d\mu_2 d\mu_1$$

idea  
curve fitting revisited  
comparing distributions



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



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

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

$$\begin{aligned} P(\mu_1 > \mu_2 | D_1, D_2, I) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\mu_1} P(\mu_1, \mu_2 | D_1, D_2, I) d\mu_1 d\mu_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\mu_1} P(\mu_1 | D_1, I) P(\mu_2 | D_2, I) d\mu_1 d\mu_2 \end{aligned}$$

we only compare means: marginalization over  $\sigma_k$

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

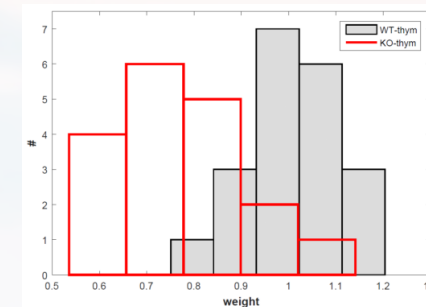
$$\sim \int_0^{\infty} P(D_k | \mu_k, \sigma_k, I) P(\mu_k, \sigma_k | I) d\sigma_k$$

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

idea  
curve fitting revisited  
comparing distributions



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





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

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

$$P(\mu_1 > \mu_2 | D_1, D_2, I) = \int_{-\infty}^{\infty} \int_0^{\mu_1} P(\mu_1, \mu_2 | D_1, D_2, I) d\mu_1 d\mu_2$$

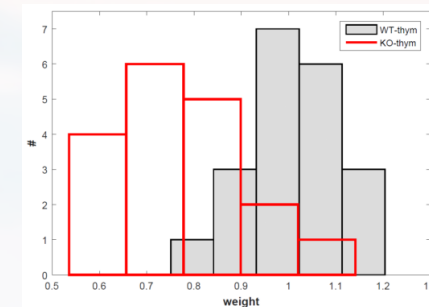
$$= \int_{-\infty}^{\infty} \int_0^{\mu_1} P(\mu_1 | D_1, I) P(\mu_2 | D_2, I) d\mu_1 d\mu_2$$

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

likelihood function  
→ the actual model

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

idea  
curve fitting revisited  
comparing distributions



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



**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:

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

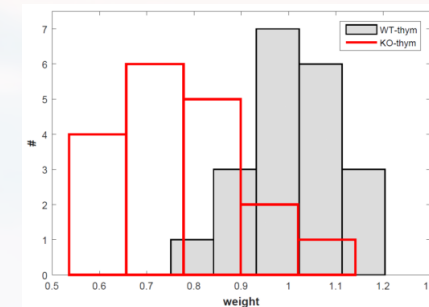
$$\hat{z} = \hat{\mu}_1 - \hat{\mu}_2$$

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

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

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

idea  
curve fitting revisited  
comparing distributions



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



**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:

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

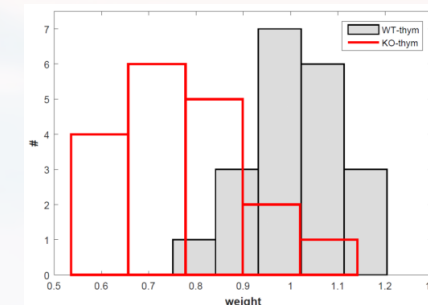
$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} x_{i,k}$$

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

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

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

idea  
curve fitting revisited  
comparing distributions



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

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

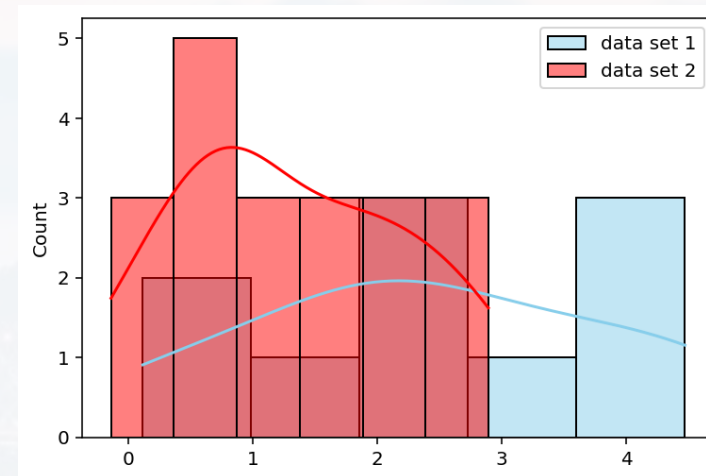
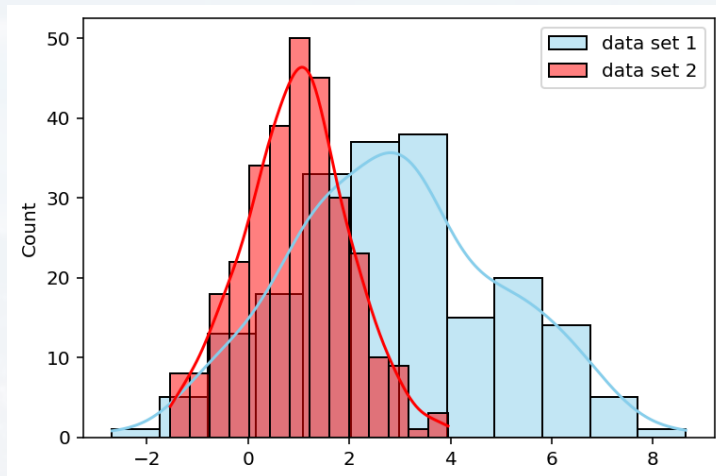
- changing  $\mu_1$  with  $\mu_2$  changes  $P(\mu_1 > \mu_2 | D_1, D_2, I)$  to  $1 - P(\mu_1 > \mu_2 | D_1, D_2, I)$



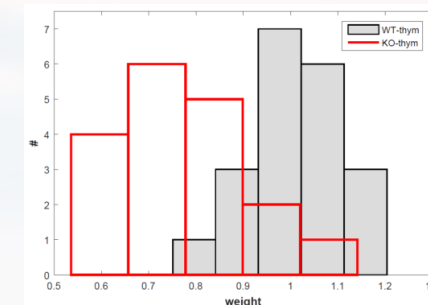


**example:**  $\mathcal{N}(y_k, \sigma_k)$

Check out: mean\_var\_analysis\_example.py



idea  
curve fitting revisited  
comparing distributions



$D_m$ :  $N_m$  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

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

M.ModelSelection()

0.711

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

M.P\_Means()

(0.99, 0.01)

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

Thank you very much for your attention!

