# Review

D. Jason Koskinen
koskinen@nbi.ku.dk

*Advanced Methods in Applied Statistics*
*Feb - Apr 2018*

Photo by Howard Jackman

# Exam Notes

- Submission is **both**:
  - A nicely written and composed PDF file devoid of code
    - You can create a latex/Word/OpenOffice/etc. template right now and save yourself time
  - The code you used to generate your results
- If you have problems email me. Worst scenario is you get a reply “I cannot help you with XXXXXX”.
- Especially for Ph.D. students,  if you don’t get an exam link via email, I will post the exam on the course webpage within a few minutes of start time CET and you can email your exam submission(s): code and PDF write-up.

# Announcements

- I will not be reviewing everything in the course today
  - Some text-heavy slides are included online, but won't be covered in class.
- An omitted topic in today's review may appear on the exam

# Likelihoods

$f()$ is commonly the probability distribution function

- The likelihood is the product of the individual probability (or probabilities for multiple parameters) of parameters (θ) which produce the observed outcomes ($x_i$)

$$\mathcal{L}(\theta) = \prod_{i=0}^{N} f(x_i; \theta)$$

- The likelihood ($\mathcal{L}$ or L) given the observed data ($x_i$) for the parameters (θ) is equal to the probability ($\mathcal{P}$) given the parameters (θ) of getting the observed data ($x_i$)

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

# Maximum Likelihood Method

- A very powerful and general method of parameter estimation when the functional form of the parent distribution is known.
- For large samples the estimators are normally (gaussian) distributed and hence the variances of the estimates are simple to determine.
- Even for small samples the estimators possess most of the expected "good" properties.

- Define: The estimate, $\hat{\lambda}$, is the value that maximizes the likelihood function.
- Since the likelihood function and the natural logarithm (ln) of the function have the same point for their maximum, we typically use the ln(L) since sums are easier to handle than products:

$$\ln L = \sum_{i=1}^{n} \ln(f(x_i; \lambda))$$

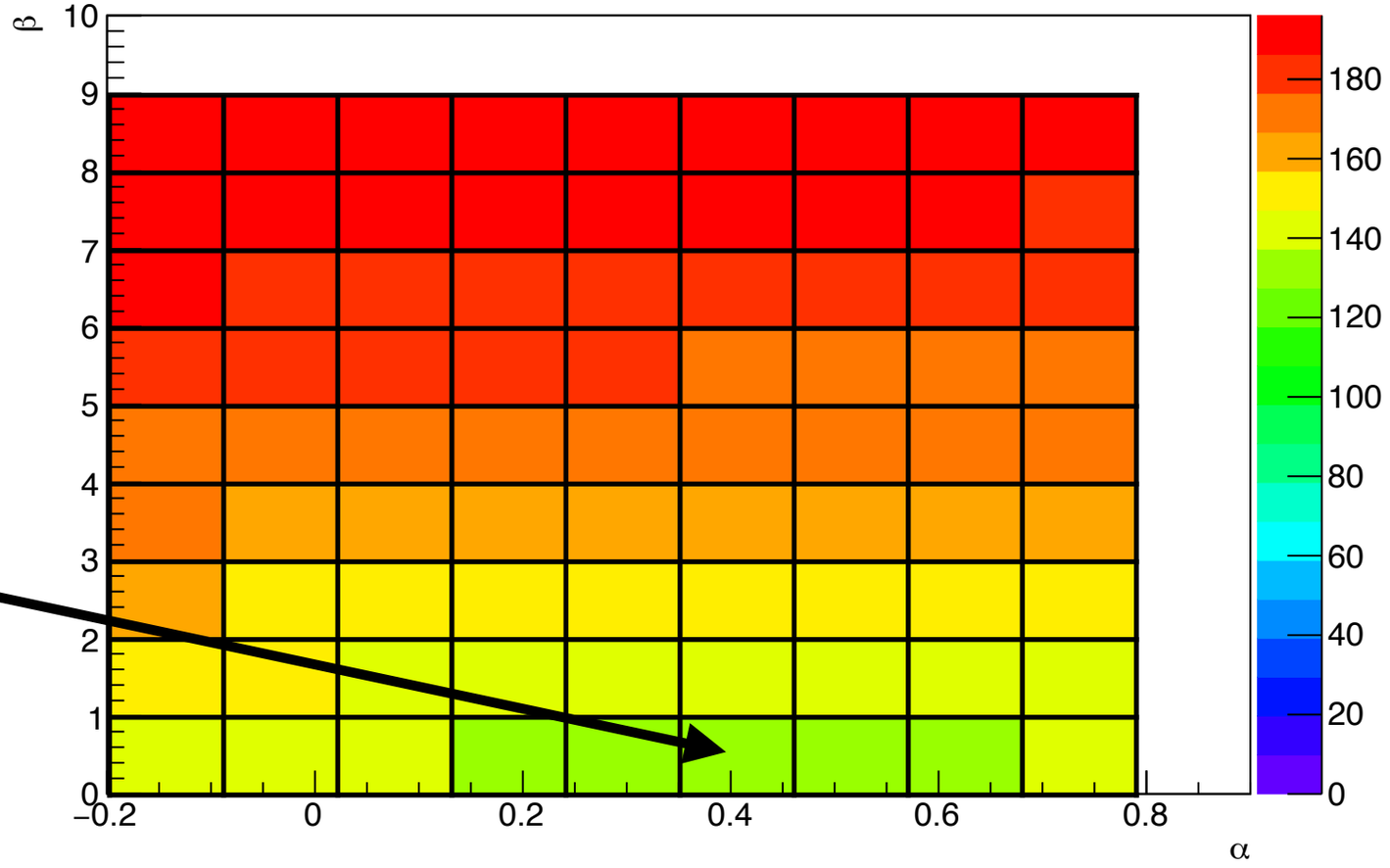
# Likelihood Minimizers

- An computation tool that finds the minima/maxima, depending on how it is setup, within a likelihood space
- Minimizers are named as such for a reason. If you want to find the maximum likelihood, you often use a minimizer and minimize over in terms of the negative ln(likelihood), -LLH, or -2*LLH
- Minimizers are often accurate, but can be sensitive to tuning parameters and local minima/maxima
- A good first step is to run a coarse scan over the -LLH space to search out good regions to start the minimizer

# Raster Scan

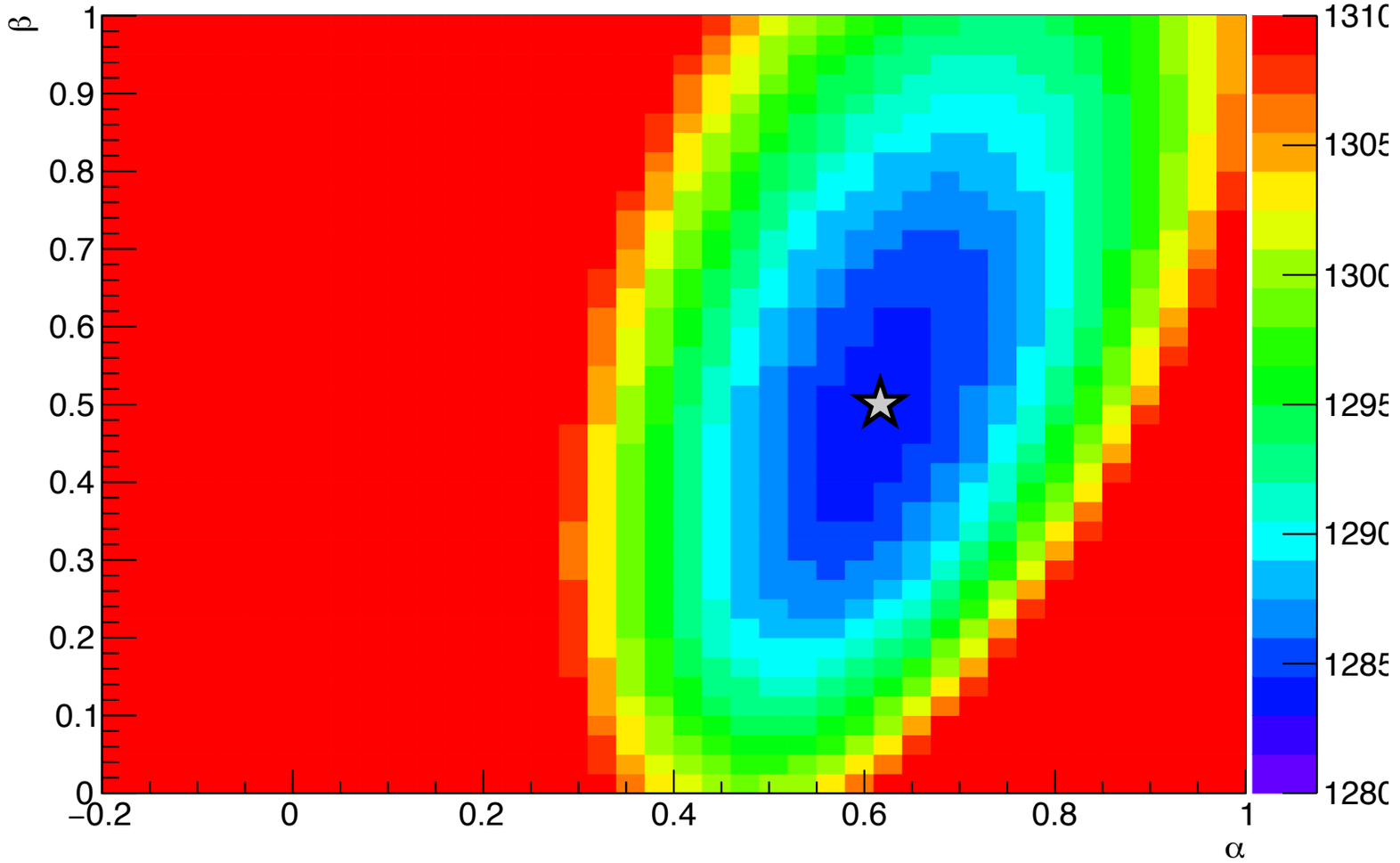
- This is a semi-coarse sampling of the LLH space. Establish which region(s) of the scanned  parameter values have the best LLH and start your fit there, or at multiple points near the best LLH.

Start somewhere around here



# Exercise 3 cont.

- Likelihood landscapes are important to visualize and understand… super important. Plot them whenever possible to understand the topology that your minimizer encounters
- For values of $\alpha$=0.6 and $\beta$=0.5 for the previous formula/PDF make a 2D plot of the likelihood or LLH landscape




Zoomed in

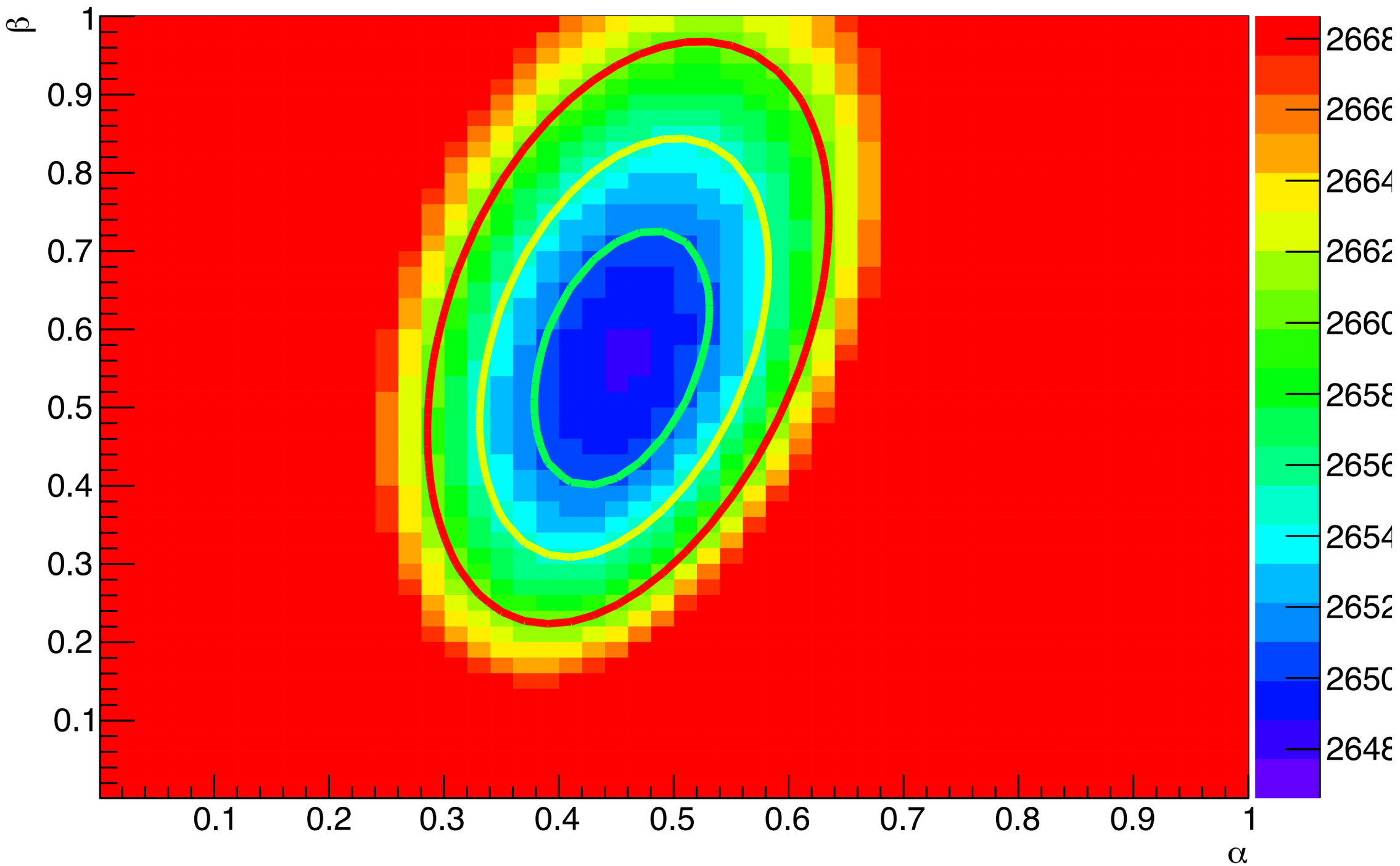
# ln(Likelihood) and 2*LLH

- A change of 1 standard deviation (σ) in the maximum likelihood estimator (MLE) of the parameter θ leads to a decrease in the in the ln(likelihood) of 1/2 for a gaussian distributed estimator
  - Even for a non-gaussian MLE, the 1σ region defined as LLH-1/2 is a good approximation
  - Because the regions defined with ΔLLH=1/2 are consistent with common $\chi^2$ distributions multiplied by 1/2, we often calculate the likelihoods as 2*LLH
- Translates to >1 parameters too, with the appropriate change in 2*LLH confidence values
  - 1 parameter, Δ(2LLH)=1 for 68.3% C.L.
  - 2 parameter, Δ(2LLH)=2.3 for 68.3% C.L.

# Variance/Uncertainty - Using LLH Values

- The LLH (or -2*LLH) landscape provides the necessary information to construct 2+ dimensional confidence intervals, provided the respective MLEs are gaussian or well-approximated as gaussian
- Some minimization programs will return the uncertainty on the parameter(s) after finding the best-fit values
  - The .migrad() call in iminuit
  - It is possible to write your own code to do this as well

# Contours on Top of the LLH Space



# Beyond Parameter Estimation

- Often we want to know if our model fits the data, or vice versa, where we find ourselves in the realm of wanting to test one hypothesis against another
  - Is my event signal or background?
  - In comparison to model $H_1$ can an alternate model $H_0$ be excluded as incompatible with the data?

# Maximum Likelihood Ratio

- An very common test-statistic for the likelihood ratio is:

$$\Lambda(\theta, x_{obs}) = -2 \ln \frac{\mathcal{L}(\theta_0 | x_{obs})}{\mathcal{L}(\hat{\theta} | x_{obs})}$$

  - Difference between the null hypothesis in the numerator and the alternative hypothesis in the denominator is that the null hypothesis has a **fixed value** of one (or more) of the θ parameters whereas the alternative hypothesis **fits/maximizes** the parameter.
- For a normal distributed, i.e. gaussian, variable the ratio follows the $\chi^2$ distribution,
  - $N_{DOF}$ = difference in dimensionality between the models
  - Also requires that Wilk's Theorem is satisfied

# Wilk's Theorem… Kinda

- As the number of data points approaches infinity, the LLH ratio converges to a $\chi^2$ distribution if $H_0$ is true

$$\Lambda(\theta, x_{obs}) = -2 \ln \frac{\mathcal{L}(\theta_0 | x_{obs})}{\mathcal{L}(\hat{\theta} | x_{obs})}$$

- But there are regions where the gaussian, and therefore Wilk's and our use of $\chi^2$, breaks down
  - **Low** number of events where the probability switches from gaussian to poisson
  - **Bounds** on the model parameters, e.g. as n→infinity the parameter does not smoothly vary, but has some truncation or discrete behavior
  - Parameters that have a **near-infinite** variance

# More Test Statistics Methods

- Use of the likelihood value, ln(likelihood), and ln(likelihood) ratio are complemented by other test-statistics
- The most common is the Kolmogorov-Smirnov test
- The KS-test is a quantitative metric of the statistical compatibility between two spectra
  - Analytic function versus data
  - data versus data
  - I guess analytic function versus analytic function too, but that's just silly

# Goodness-of-fit

- Pearson's Chi-square Test
  - A goodness of fit test that could be applied to a histogram of observed values, x, with N bins. For the number of entries in bin i, $n_i$, and the number of expected entries for the same bin, $\lambda_i$, the test statistic becomes:

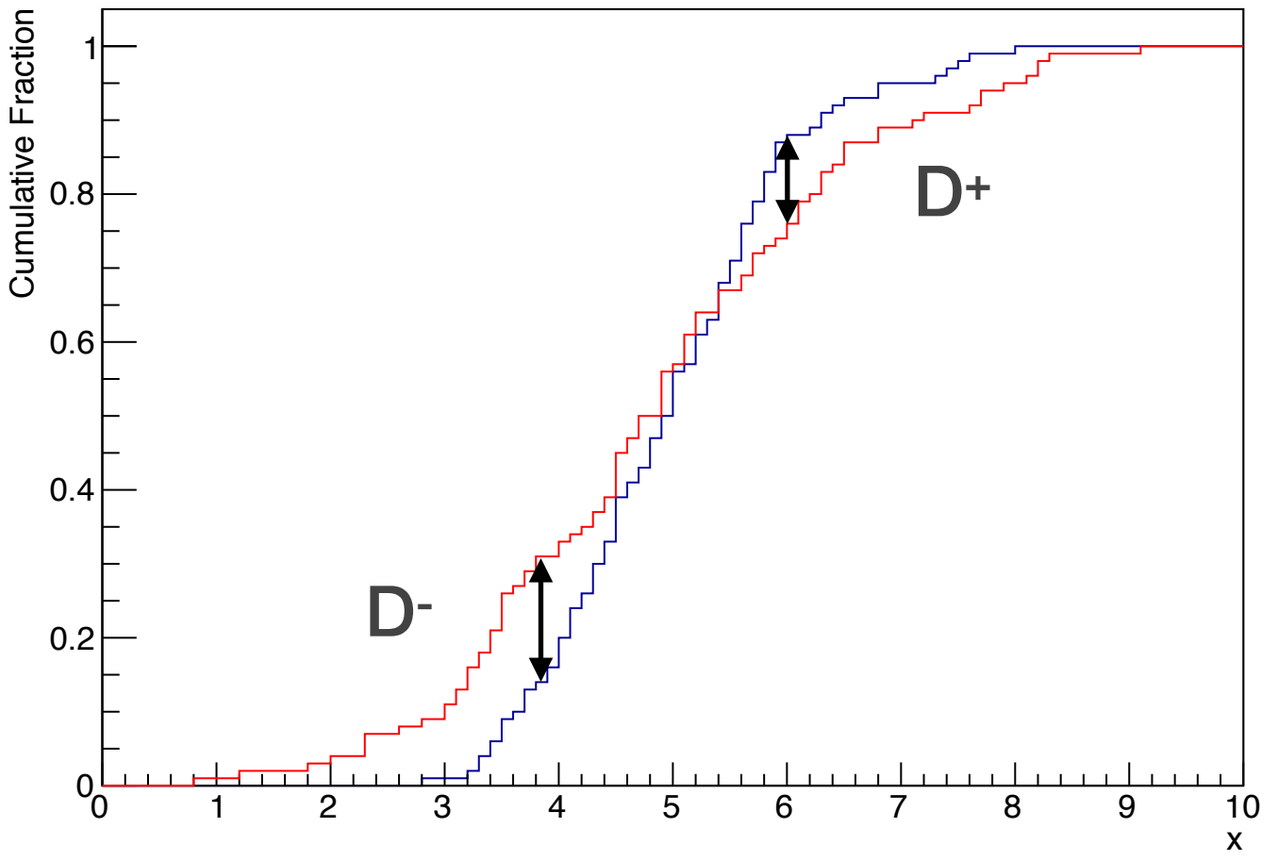
$$T = \chi^2 = \sum_{i=1}^{N} \frac{(n_i - \lambda_i)^2}{\lambda_i}$$

    - If the data are Poisson distributed, and the number of entries is not too small in each bin (>5), then T follows a chi-square distribution of N degrees of freedom. This is true regardless of the distribution of x, implying the chi-square test is distribution free.

- Even though finding the maximum likelihood estimator (MLE) best-fits are often done using an unbinned likelihood, it is often useful to use histograms to get a (reduced) chi-squared value as a goodness-of-fit parameter

# One/Two Sided Test

- Most often we want to know about the greatest difference between the two distributions/samples, regardless of whether sign (+/-) of the deviation. This is a two-sided or two-tailed test.
- A one-sided test is where we want to know about deviations in only a single direction, i.e. + or - deviations.



# Bayesian

# Transition to Bayes

- The maximum likelihood approach is both effective and powerful, but does not necessarily take into account any preferences or prior information that may produce a more informed or accurate result
- Thankfully, we have Bayes theorem and Bayesian statistics which make explicit use of prior information
- Bayesian probabilities and statistics can encode an amount of belief in (data, model, systematics, hypothesis, parameters, etc.)

# Bayes' Theorem

- We have Bayes' theorem

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

- or sometimes

(Discrete)

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

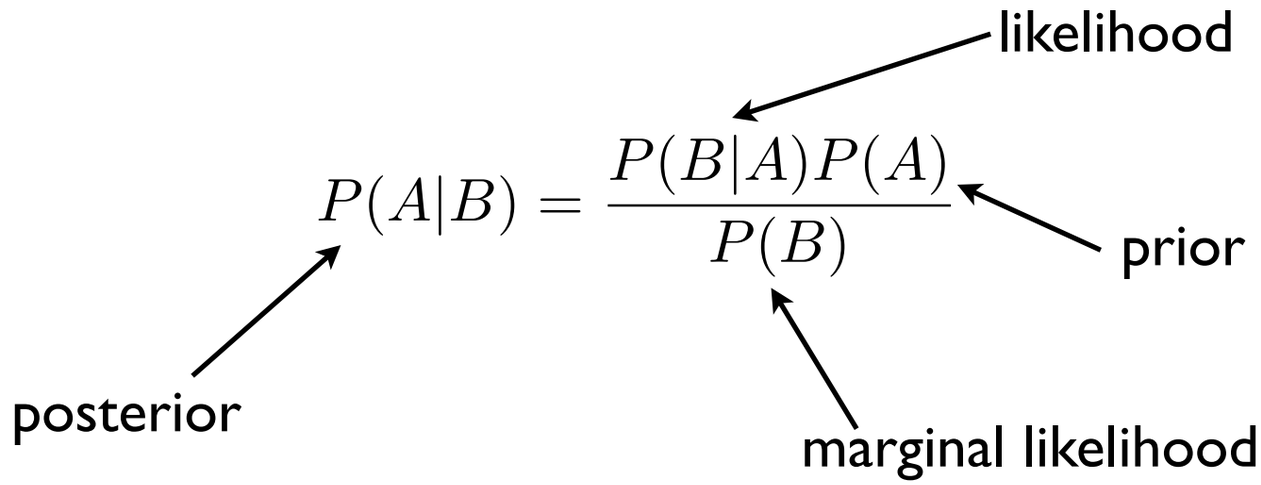
(Continuous)

$$\frac{P(B|A)P(A)}{\int P(B|A)P(A)dA}$$

- Let B be the observed data and A be the model/theory parameters, then we often want the P(A|B); the posterior probability distribution conditional on having observed B.

# Bayes' Theorem

- One can solve the respective conditional probability equations for P(A and B) and P(B and A), setting them equal to give Bayes' theorem:




$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- In the previous class we avoided dealing with the marginal likelihood, i.e. the normalizing constant, because it does not depend on the parameter(s) A. But it is an important value in order to get an accurate posterior distribution which is a useable probability.

# Application Overview

*M.A. Thomson

★ **Apply Bayes' theory to our the measurement of a parameter $x$**

- **We determine $P(\text{data};x)$ , i.e. the likelihood function**
- **We want $P(x;\text{data})$ , i.e. the PDF for $x$ in the light of the data**
- **Bayes' theory gives:**

$$P(x;\text{data}) = \frac{P(\text{data};x)P(x)}{P(\text{data})}$$

$P(\text{data};x)$ **the likelihood function, i.e. what we measure**

$P(x;\text{data})$ **the posterior PDF for $x$, i.e. in the light of the data**

$P(\text{data})$ { **prior probability of the data. Since this doesn't depend on $x$ it is essentially a normalisation constant**

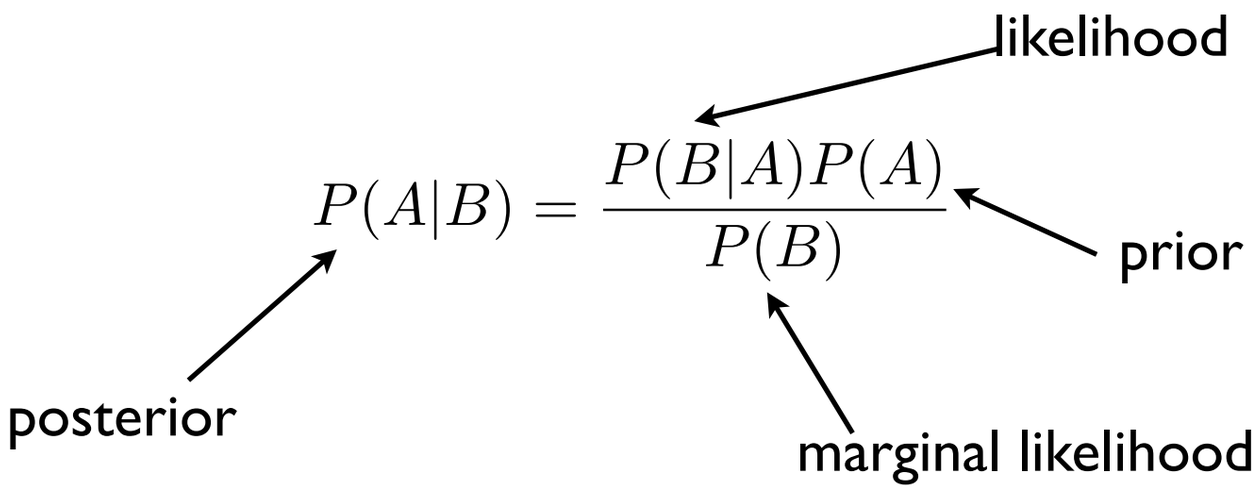
$P(x)$ { **prior probability of $x$, i.e. encompassing our knowledge of $x$ before the measurement**

★ **Bayes' theory tells us how to modify our knowledge of $x$ in the light of new data**

**Bayes' theory is the formal basis of Statistical Inference**

# Bayes for Parameter Estimation

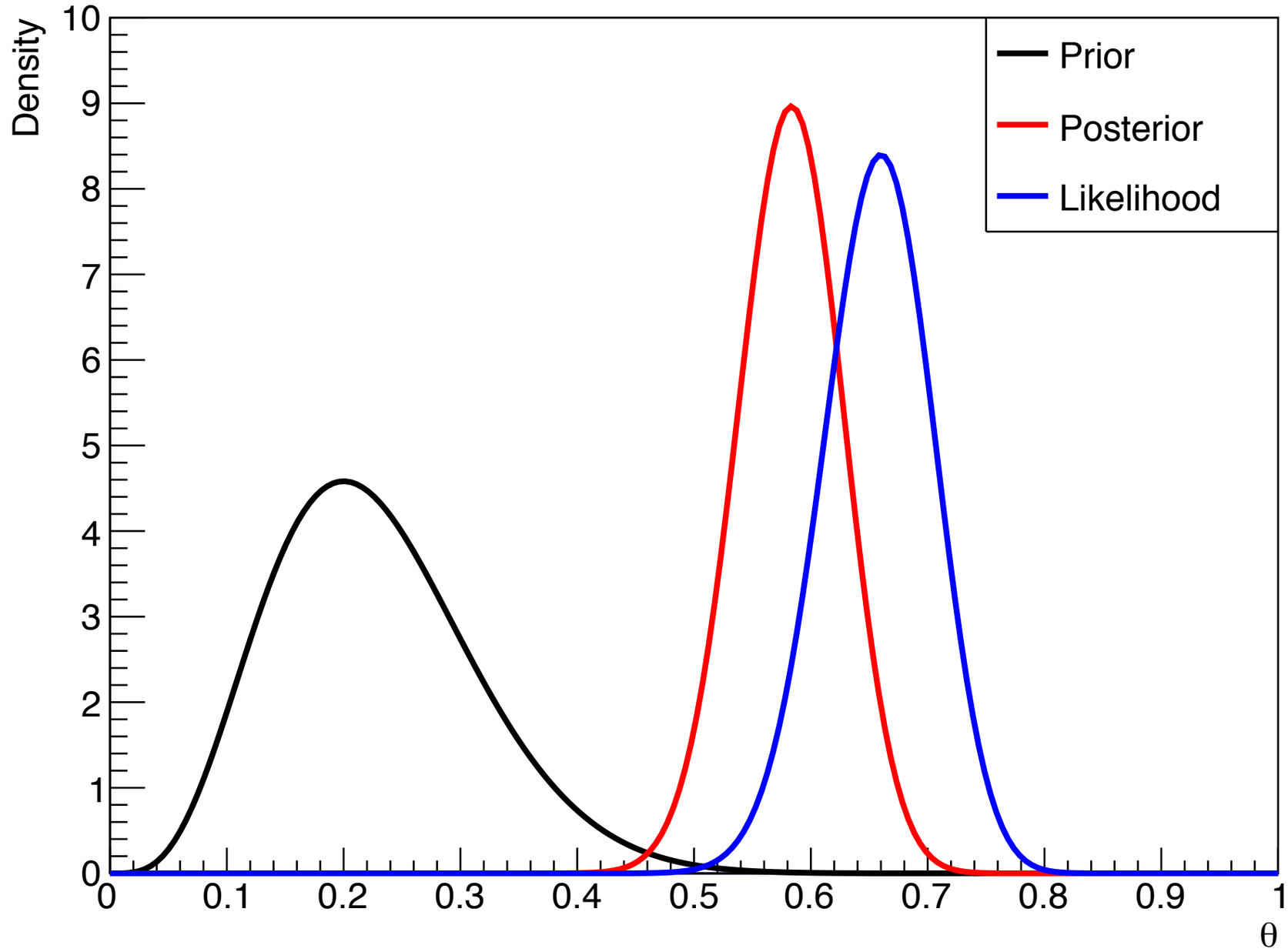
- We apply prior information not just for discrete probabilities, but for probability distributions as well
- Remember that for Bayesian analyses we include all possible values of the parameter, i.e. $\theta$
  - This means for the PDF, it will **not** be calculated at a single value of $\theta$, but over a suitable range




$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

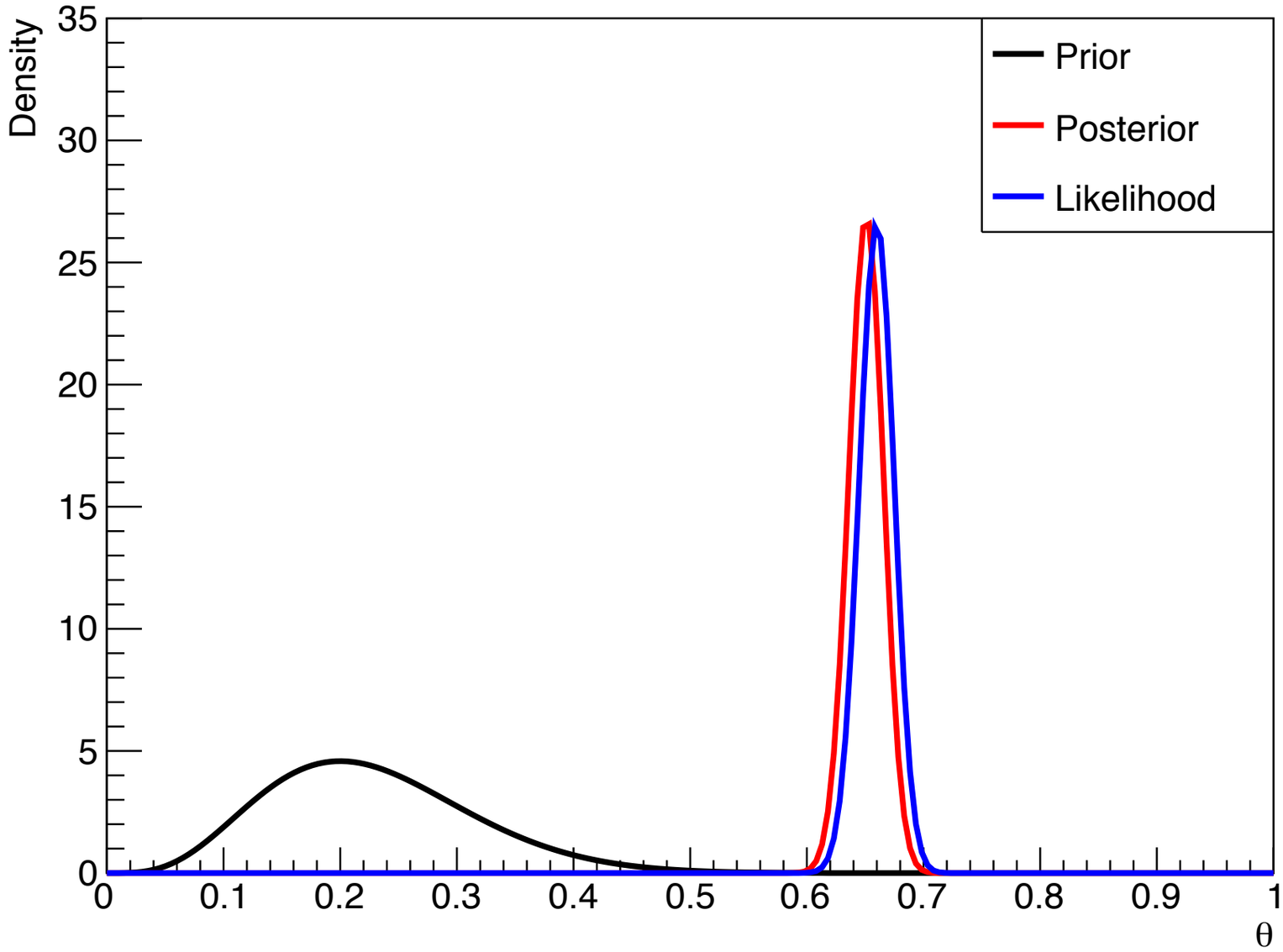
# Exercise #1 plot (from Lec. 6)

- Coin flipping bias with n throws/flips, but now in Bayesian style where we want the prob. of coming up heads (θ)



# Exercise #1 (cont.)

- With 10x more statistics, an obvious feature pops up, i.e. that as n→infinity the maximum a posterior (MAP) approaches the maximum likelihood estimator (MLE)



# Numerical Limitations

- The previous example had only 1 parameter ($\theta$) and 1 prior. When dealing with more parameters, the computational load approx. increases exponentially with the number of parameters.
  - For summation, or integration via Monte Carlo sampling, the number of points (n) grows as $\mathcal{O}(n^d)$ if n points are used to cover each parameter (d)
- It's possible to tune the number of scan or Monte Carlo points, but then the number of points necessary for calculation is the product of the number of points:

$$\prod_{i=1}^{d} n_i$$

# Markov Chains for Bayes' Stuff

- So how does a Markov chain help with establishing Bayesian posterior distributions?
- Markov chains will asymptotically approach a stable distribution, and we can give the Markov chain a distribution that is representative of the posterior. Remember that,

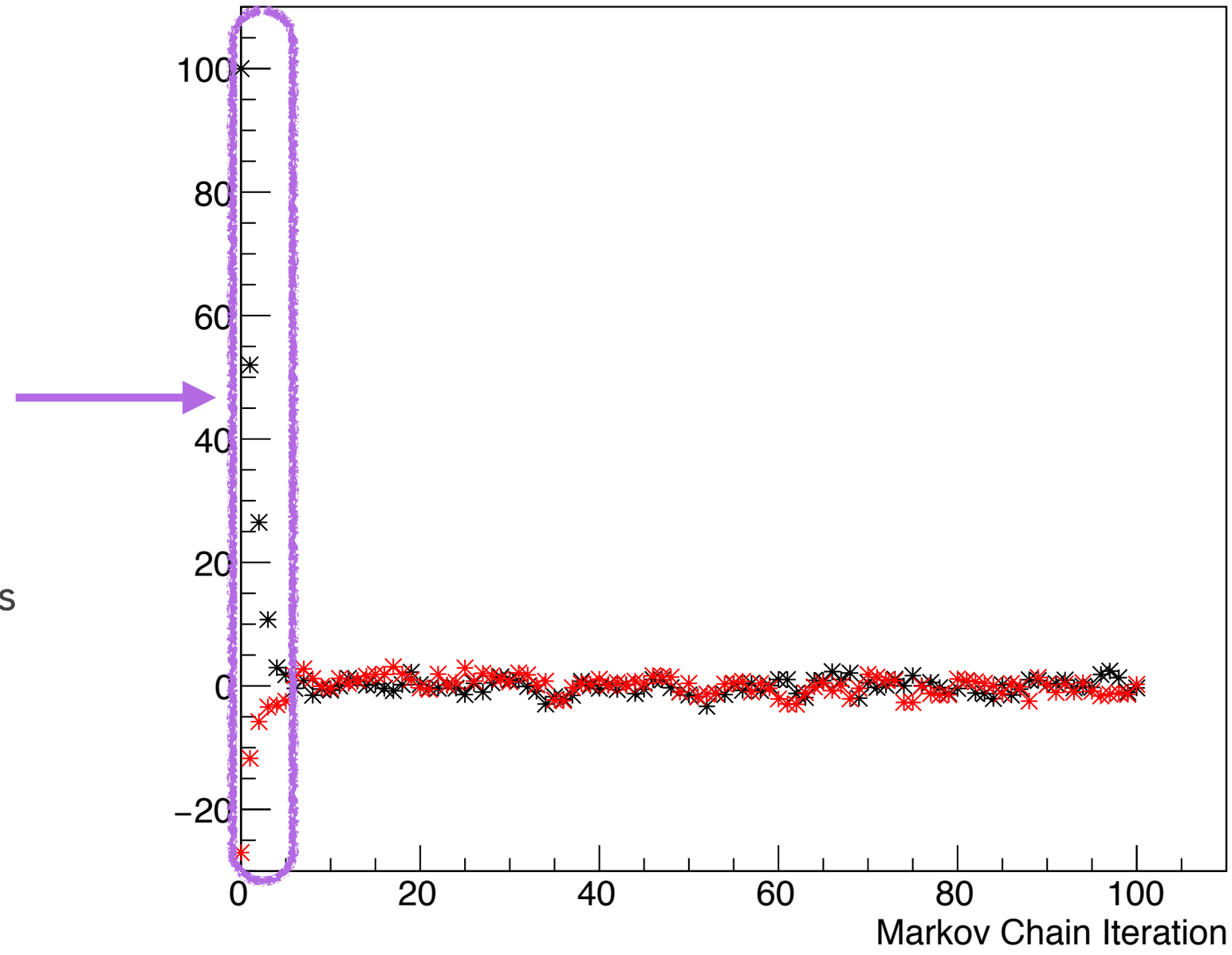
$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- So using Markov Chain Monte Carlo, the chain can start at points that are not typical of the actual posterior (which we may not know well), but after enough Monte Carlo iterations it should converge to the posterior
- Markov Chain Monte Carlo is the solution

# Exercise #2 (plots) (From Lec. 6)

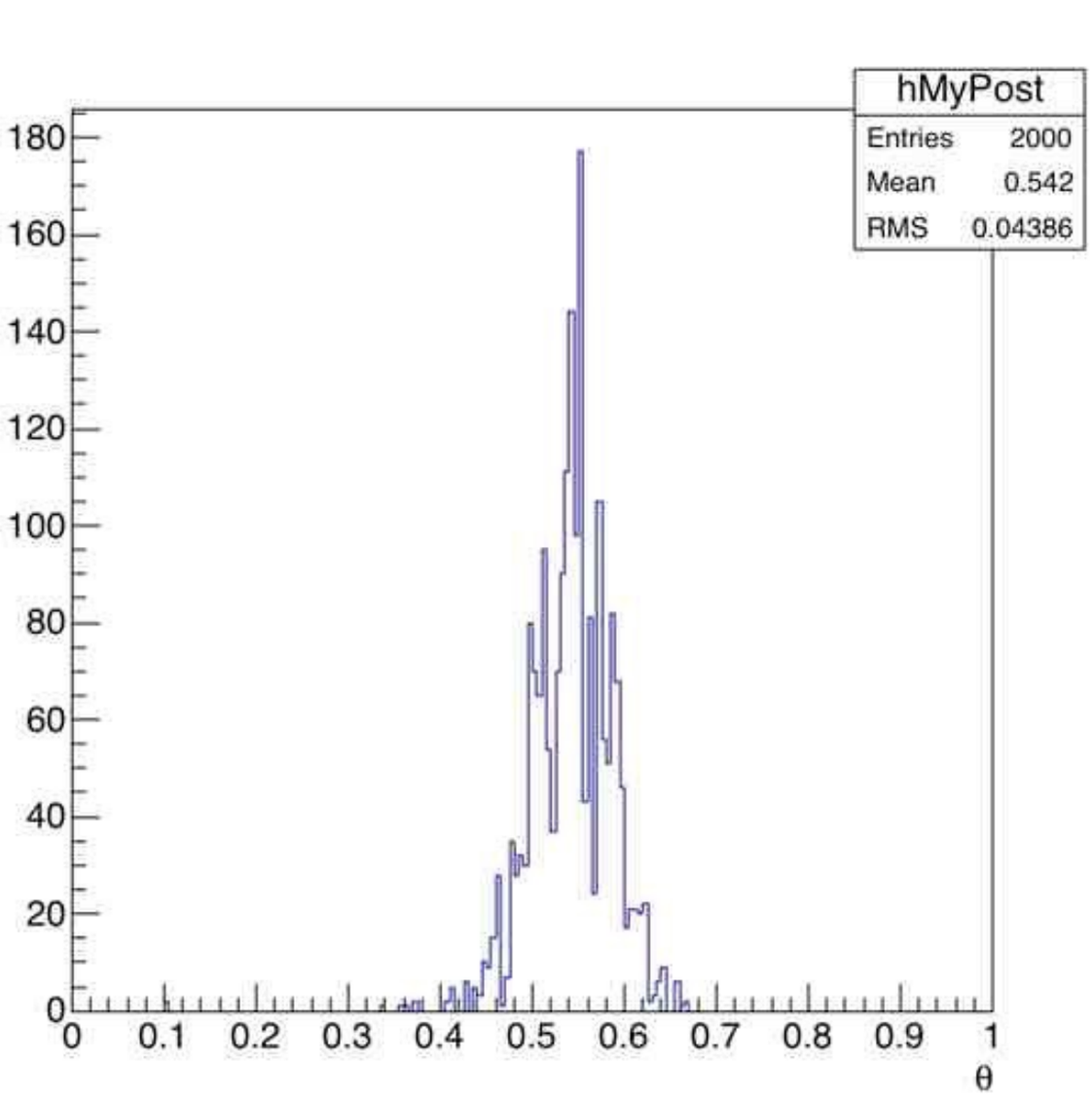
- After maybe 5-10 iterations from the starting point the chains look to converge to some stationary behavior

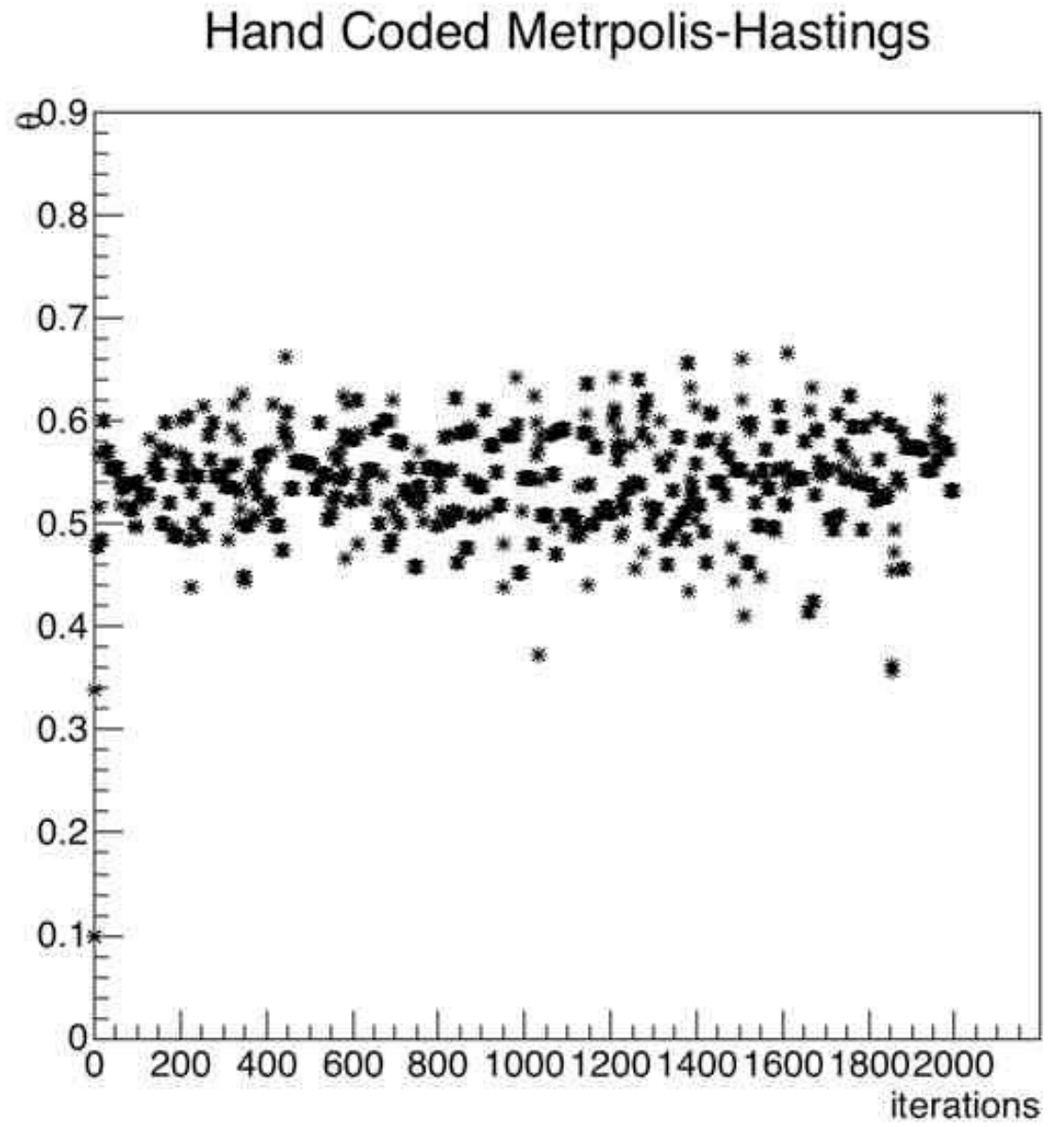
The samples before convergence are commonly known as ‘burn-in samples’ and are not often included when estimating the posterior distribution. They’re generally just discarded and understood as the cost of using Markov Chain Monte Carlos.



# Exercise #3 (cont.) (from Lec. 6)

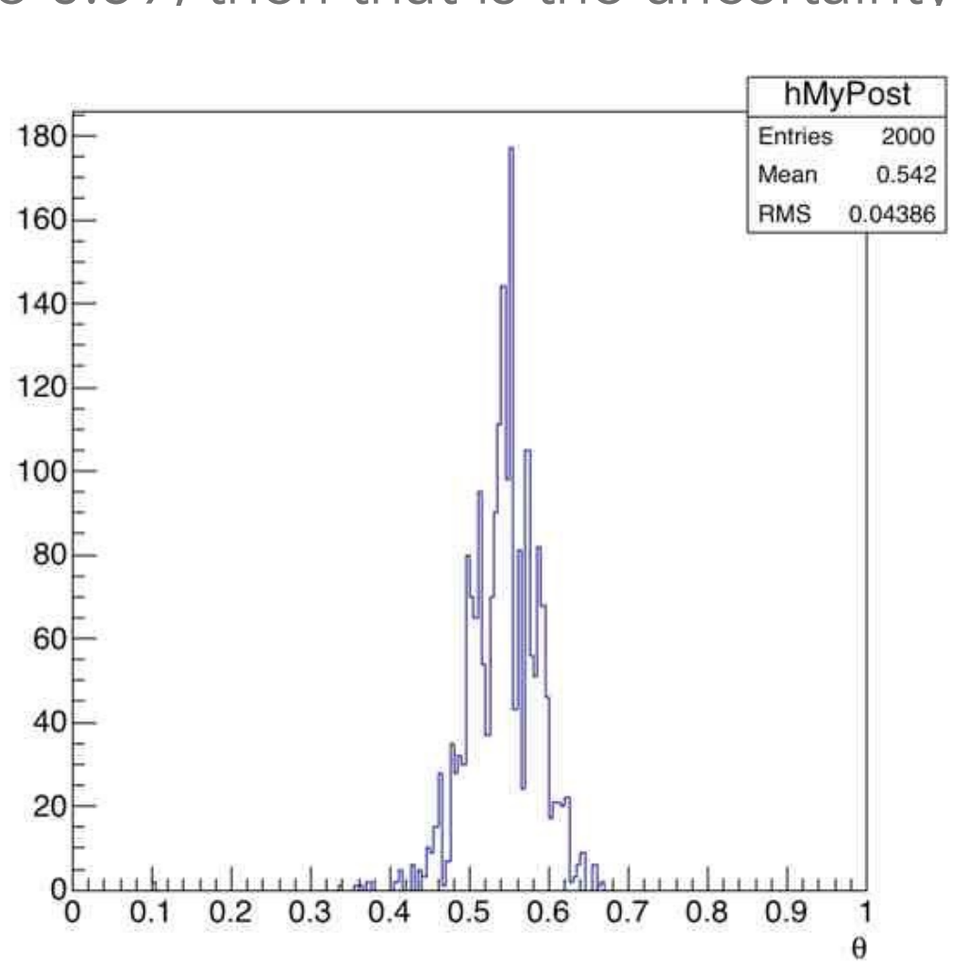
- For 2000 iterations plot Markov Chain Monte Carlo samples as a function of iteration, as well as a histogram of the samples, i.e. the posterior distribution.

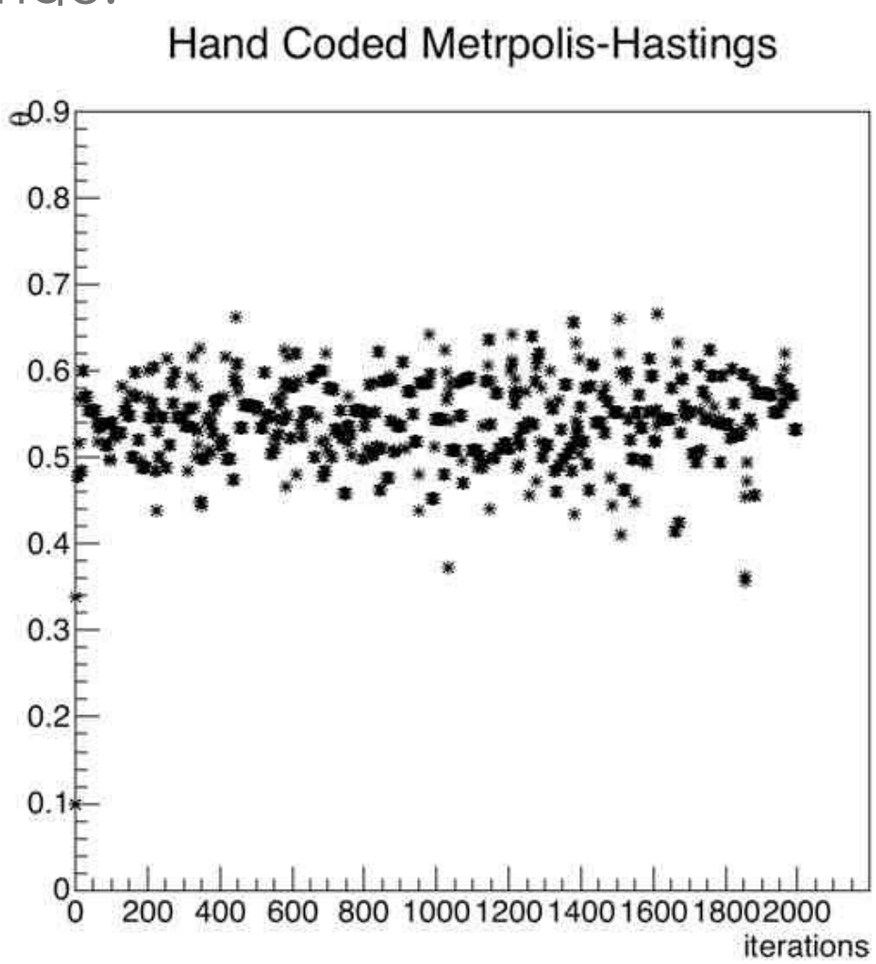


# Why the Posterior?

- The posterior distribution in the Bayesian framework provides not only the most likely value of our parameter of interest, i.e. the **maximum a posterior value**, but also the **uncertainty**. The width of the posterior gives the parameter uncertainty.
- For the example below, if 68.3% of the posterior MCMC iterations occur from 0.5 to 0.59, then that is the uncertainty range.



# Bayesian Complication

- Unlike the maximum likelihood approach, where we normally just have to know the -2*LLH value which can be converted to a probability, the Bayesian approach can be more resource intensive
- In order to get a 5σ confidence limit, we need approx. 1.7M stable posterior points/iterations

# Smoothing, Interpolating, and Estimation

-

# Splines and Kernel Density Estimation

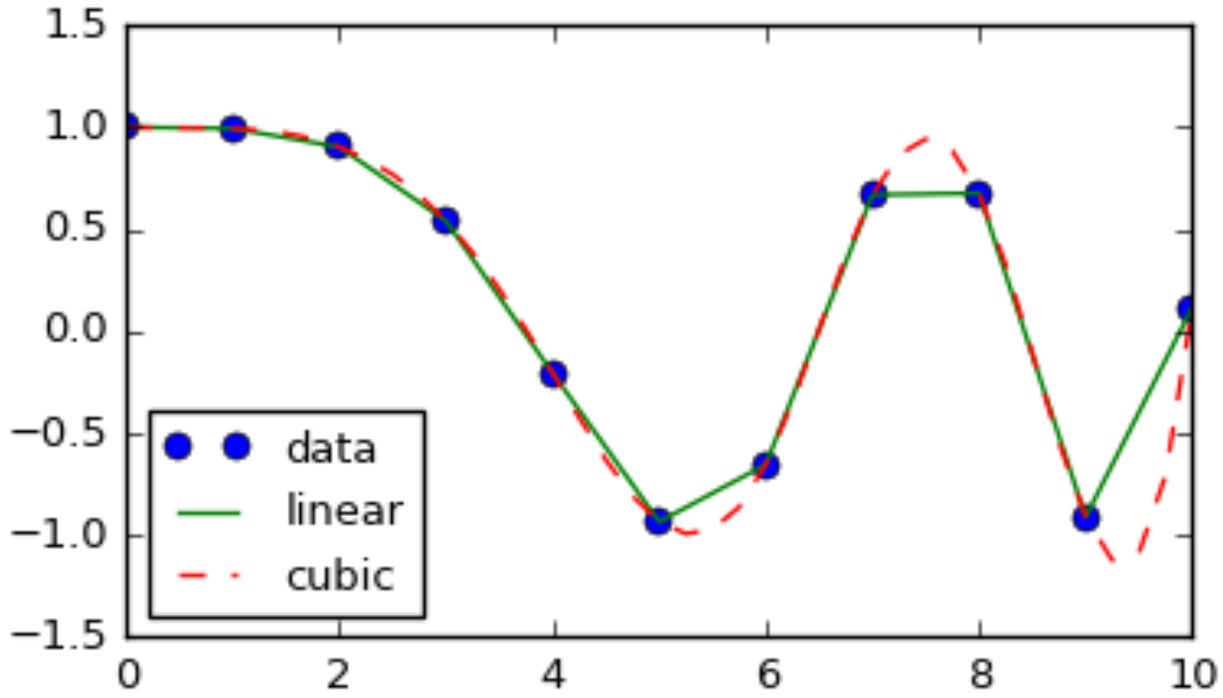
# Spline/Interpolation Use

- Where do we want to use splines?
- Computer aided drawing and graphics
- Creating continuous functions from discrete data
- Creating smooth functions from jagged or irregular data

# Common Spline Types

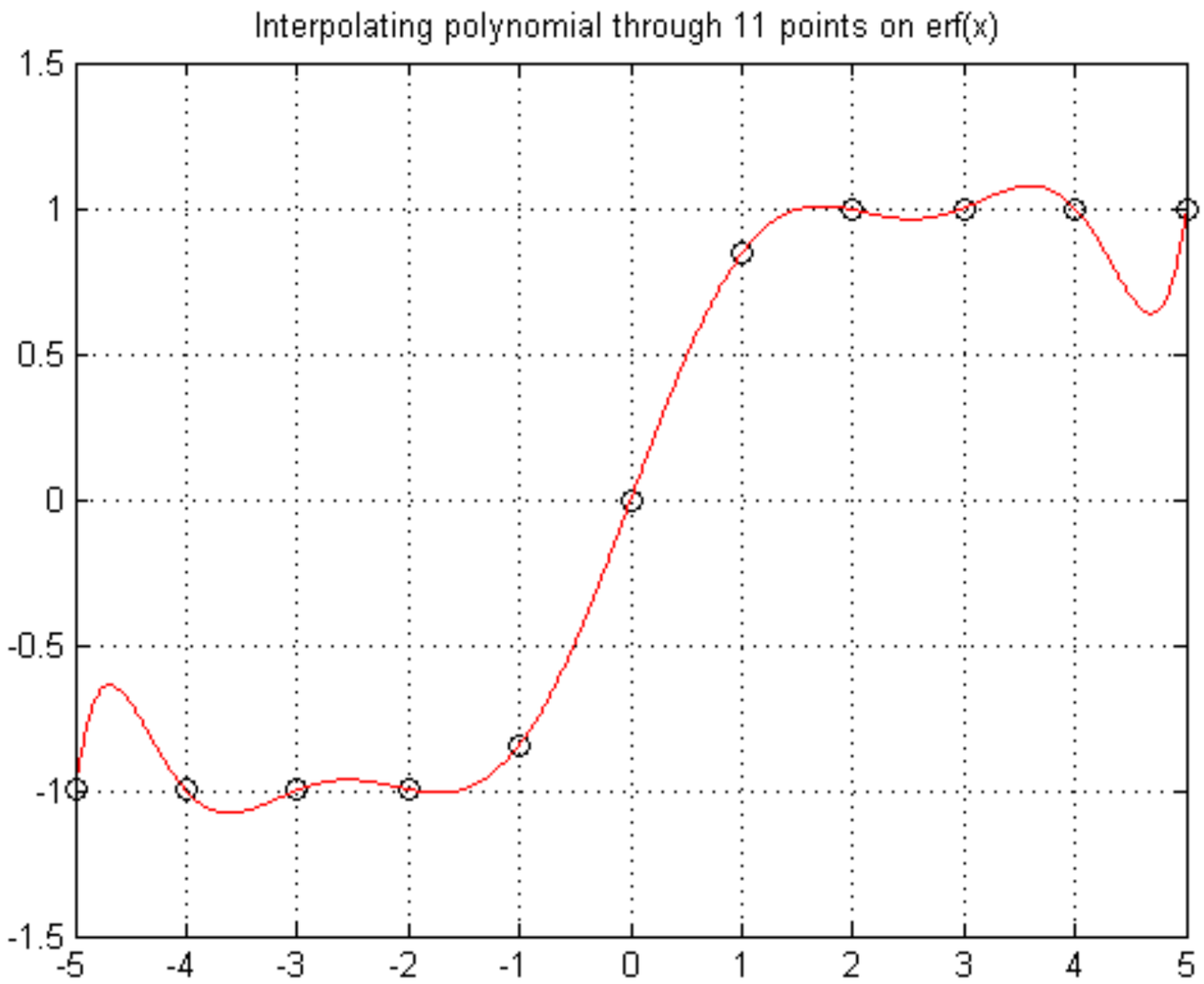
*Scipy interpolate

- Linear splines are continuous across the data points, but do not match the 1st or 2nd derivative at the knots
- Quadratic splines (not shown) match the 1st derivative but not necessarily the 2nd
- Cubic splines are continuous and match the 1st and 2nd derivative at the knots
- Hermite splines - Continuous cubic splines matching the 1st derivative but not necessarily the 2nd



# Polynomial Interpolation

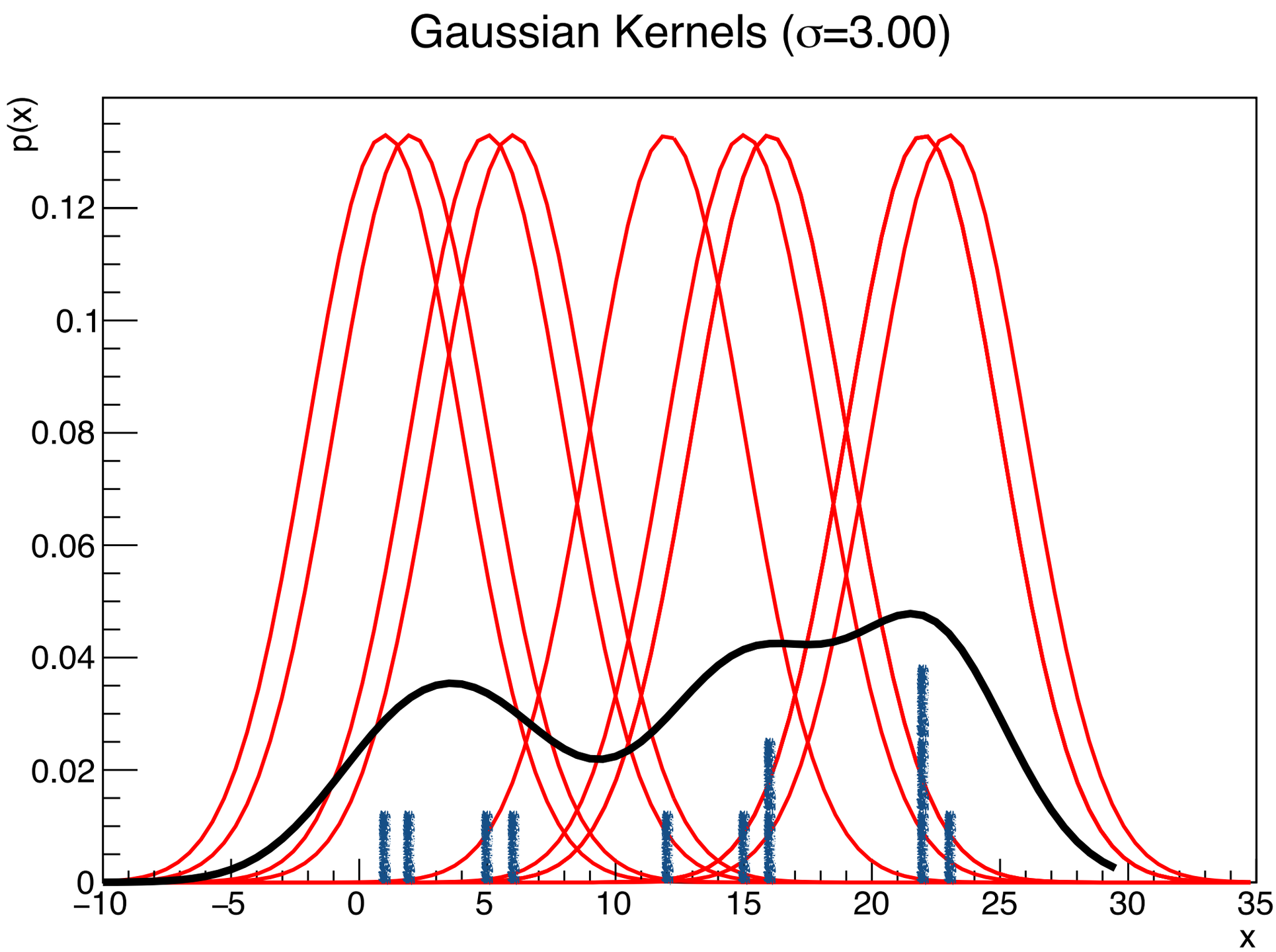
- A problem referred to as ‘ringing’ is pronounced in polynomial interpolations.



# b-splines and Smoothing

- Basis splines (b-splines) are common too. They are piecewise polynomials of order k (k=3 for cubic), where the interpolated value and most often the derivative and 2nd derivative match the adjacent piece-wise polynomials at the knots.
- There is a parameter ‘smoothness’ which can regulate the behavior of the spline
  - Large smoothness means a cubic spline is more smooth (less bumpy) , but also not constrained to go through the knots
  - Small smoothness means the splines are constrained to be close to the knots.
- Like Hermite splines, b-splines do not frequently suffer from ‘ringing’ effects

# Data Driven Density Estimation




Data = [1,2,5,6,12,15,16,16,22,22,22,23]

# Kernel Density Estimator

- The generic KDE expression can be expressed as:

$$P_{KDE}(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N} K(\vec{x})$$

- A gaussian kernel is:

$$K(\vec{x},\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^D} e^{-\frac{||\vec{x}-\vec{x}_n||^2}{2\sigma^2}}$$

- The kernel at each data point contributes a non-zero probability from [-∞,+∞] smoothly with decreasing weight as a function of distance
  - Each data point and corresponding kernel integrate to 1 over the whole parameter space

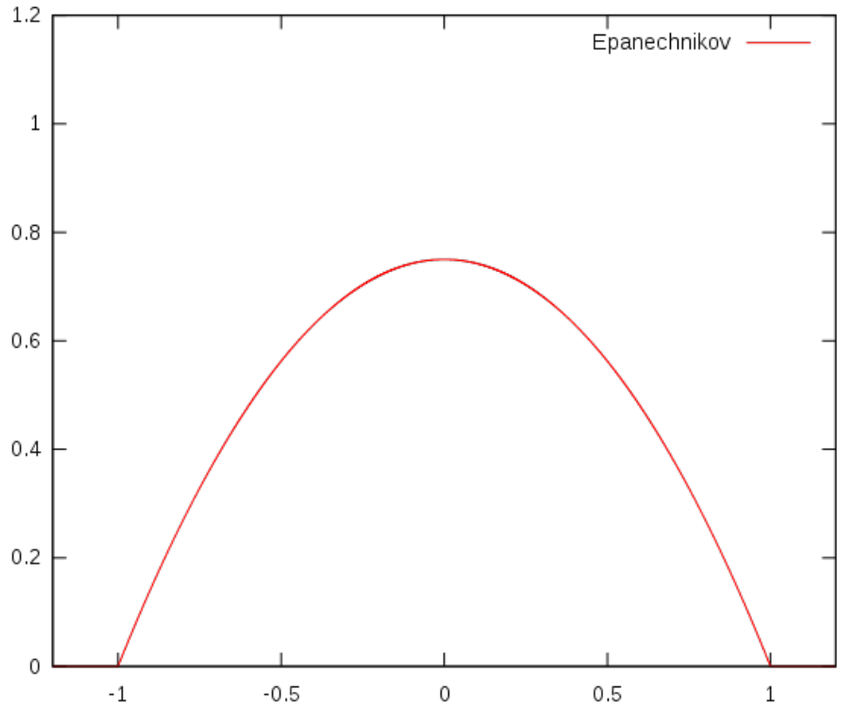
# Comment on KDE Normalizations

$$P_{KDE}(\vec{x}) = \frac{1}{N}\sum_{n=1}^{N} K(\vec{x}) \qquad K(\vec{x},\sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^D} e^{-\frac{||\vec{x}-\vec{x}_n||^2}{2\sigma^2}}$$

- The 1/N normalizes the KDE for the number of events, and an additional factor is necessary for normalizing the D-dimensional hyper-volume
- Neither normalization term in this kernel choice depend on values of $\vec{x}$

# Compact Kernel

- The gaussian kernel contributes across the whole space (infinite support), but sometimes we want compact support, i.e. zero outside of a specific range
  - Maybe some parameters are constrained to be non-negative
  - We know the physical system has either boundaries or effective cut-offs
- A common compact support kernel is the Epanechnikov kernel

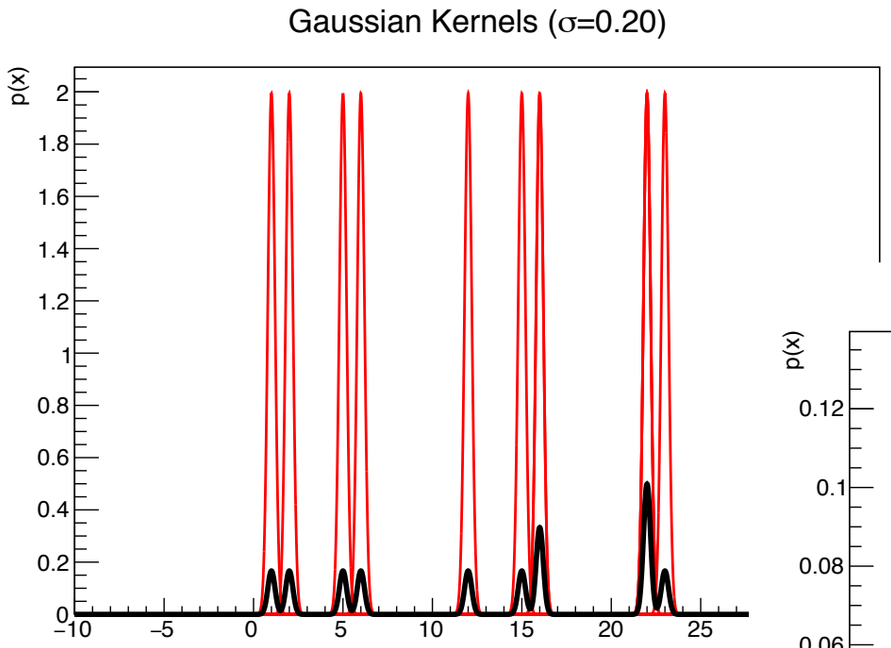


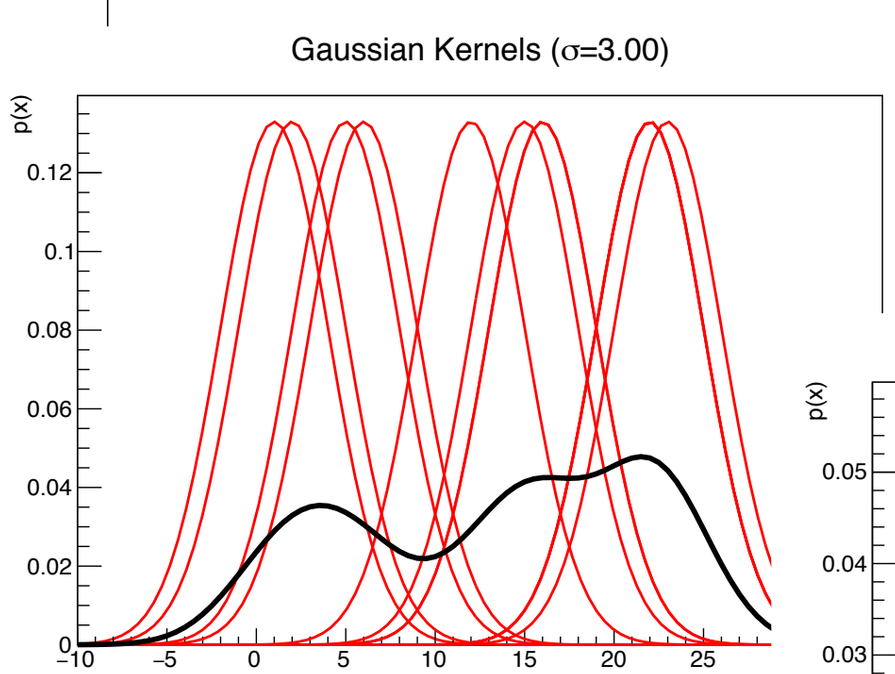

$$K(u) = \begin{cases} \frac{3}{4}(1-u^2) & \text{for } |u| \leq 1 \\ 0 & \text{for } |u| > 1 \end{cases}$$

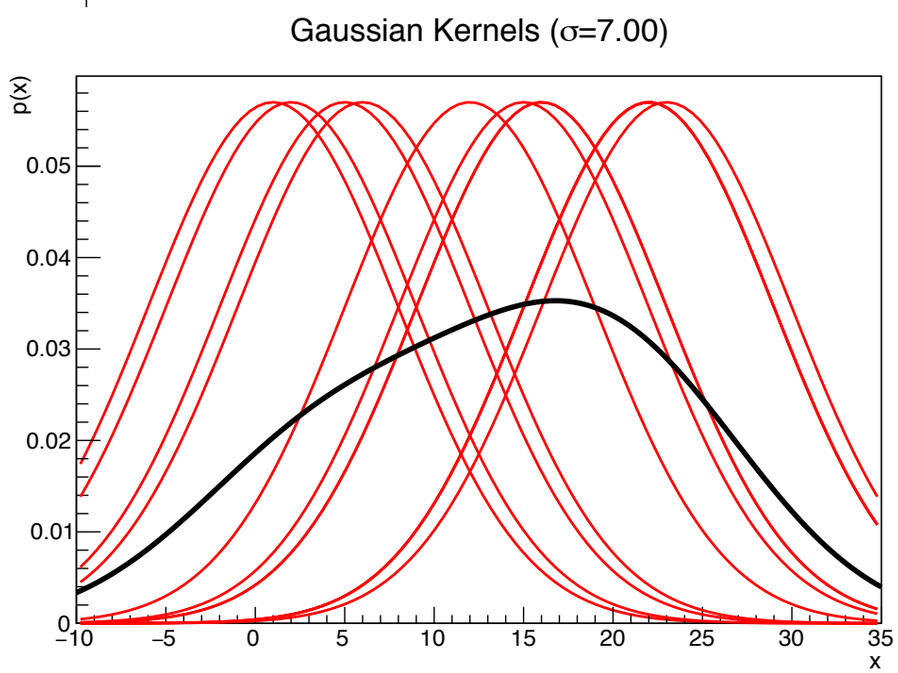
*https://en.wikipedia.org/wiki/Kernel_(statistics)

# Kernel Bandwidth

- Every KDE  is, unfortunately, strongly influenced by the kernel bandwidth, which is a user defined free parameter



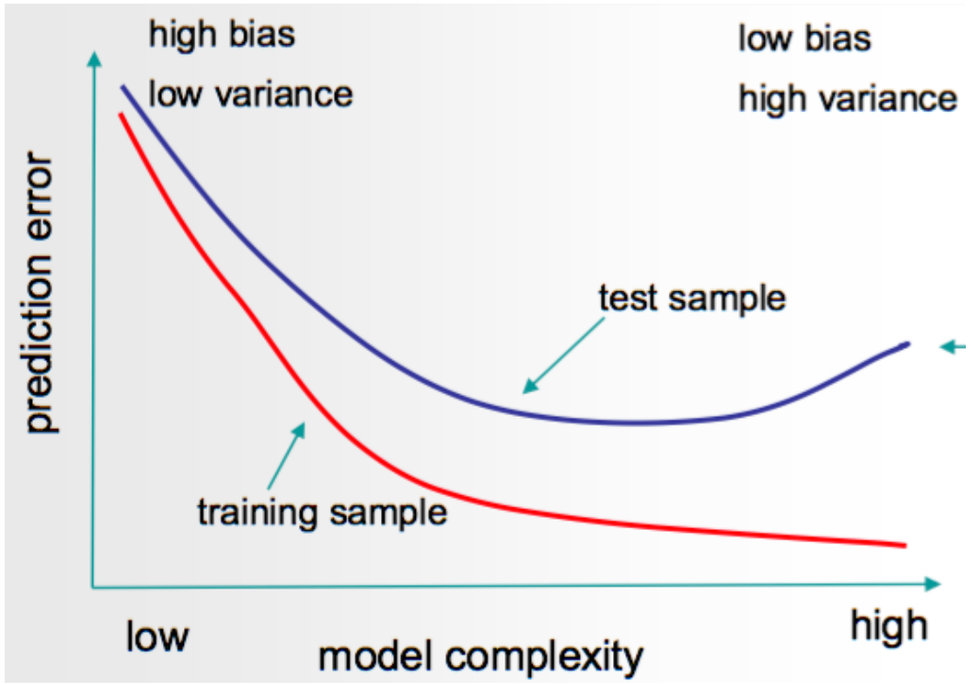
# Multivariate Method and Boosted Decision Tree

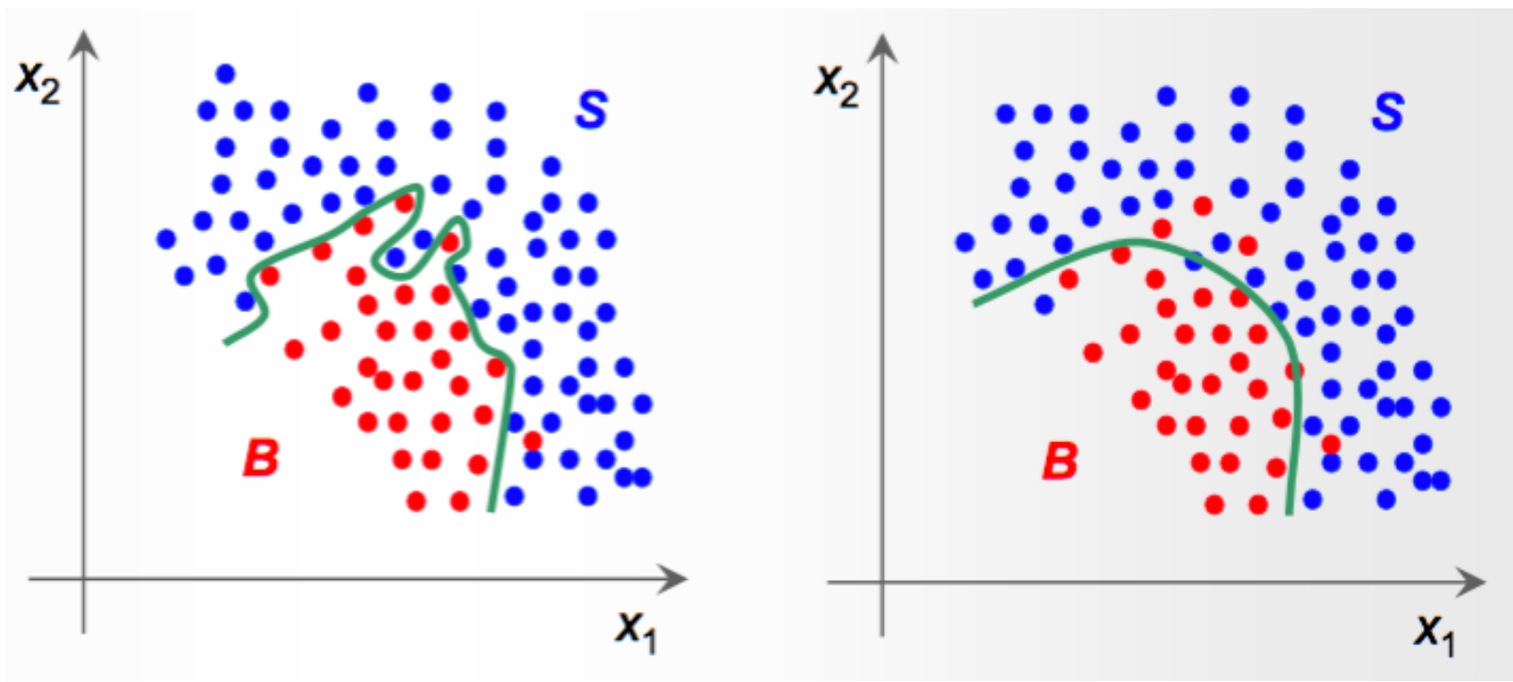
# "Simple" Problems

- Using likelihoods to separate background from signal is not always feasible
  - Likelihood may be too complicated for analytic or Monte Carlo evaluation
  - High dimensionality makes Monte Carlo computationally expensive
- Data sets which are linearly separable in variables, e.g. between signal and background, have useful tools for doing such a separation (Fisher Discriminant)
- For linear and non-linear classification scenarios and/or where the available separators are weak, there is a class of multivariate tools
  - k-Nearest Neighbor
  - Random Forest
  - Artificial Neural Networks
  - Support Vector Machine (can be a linear regression classifier too)

# Overtraining

- Machine Learning algorithms can be overly optimized wherein statistical fluctuations from the training data are wrongly characterized as true features of the distributions
  - Deficit of training data statistics versus number of variables or complexity
  - Model flexibility, e.g. many free parameters

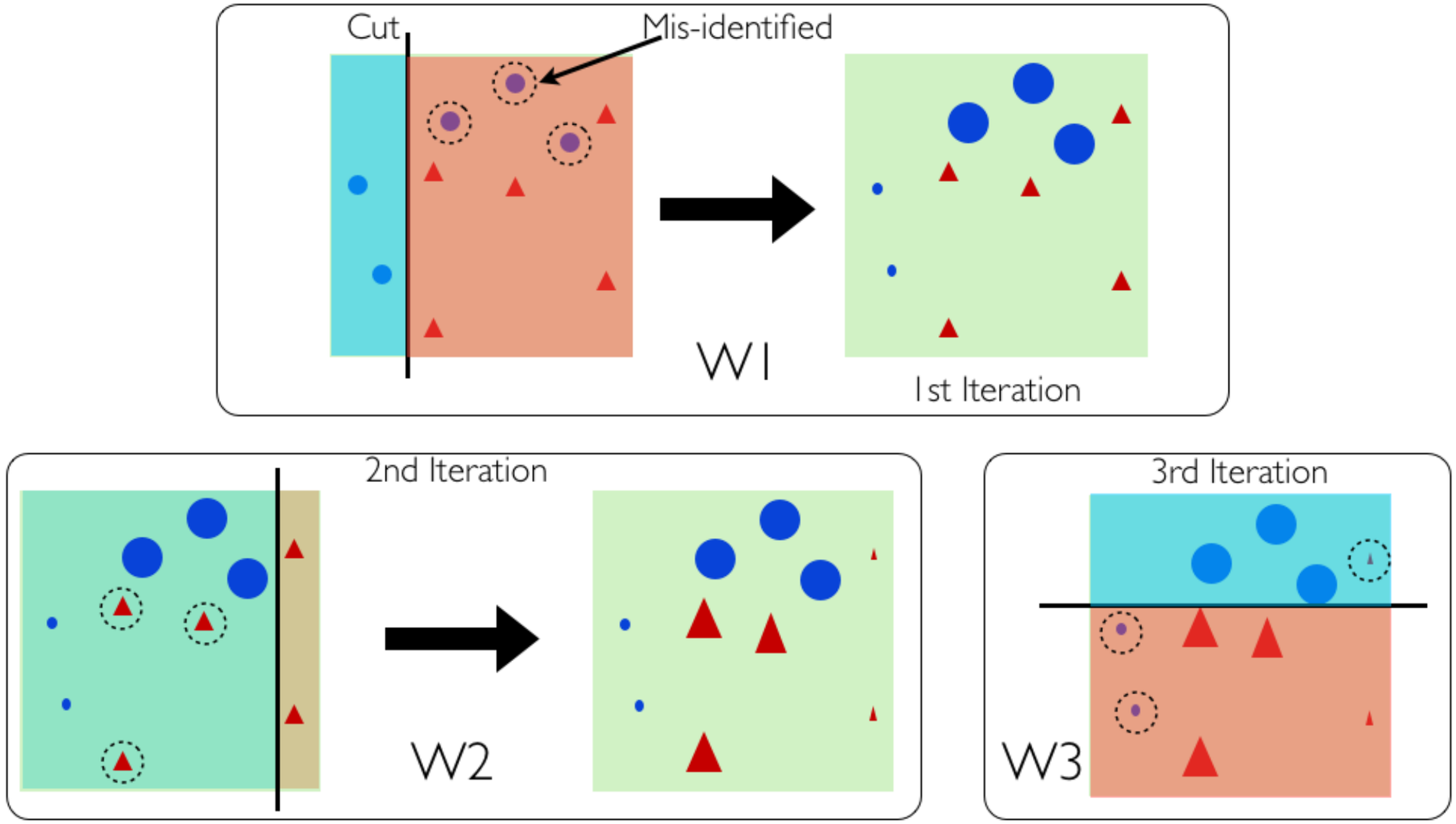



*H. Voss (MPIK)

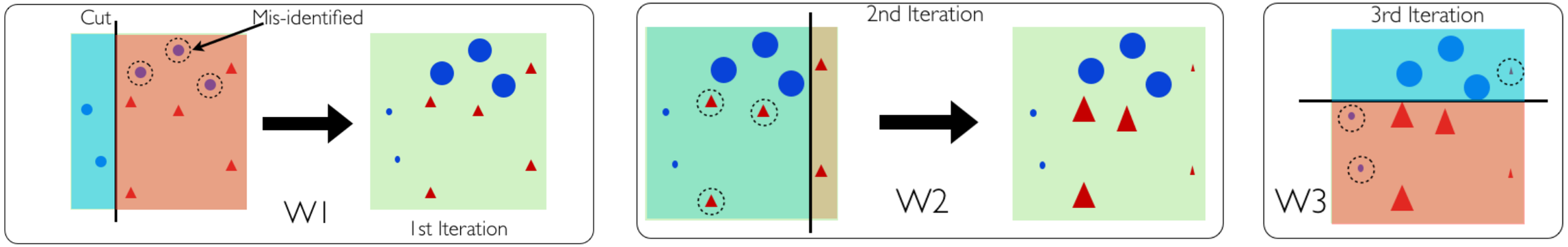
# Boosted Decision Trees

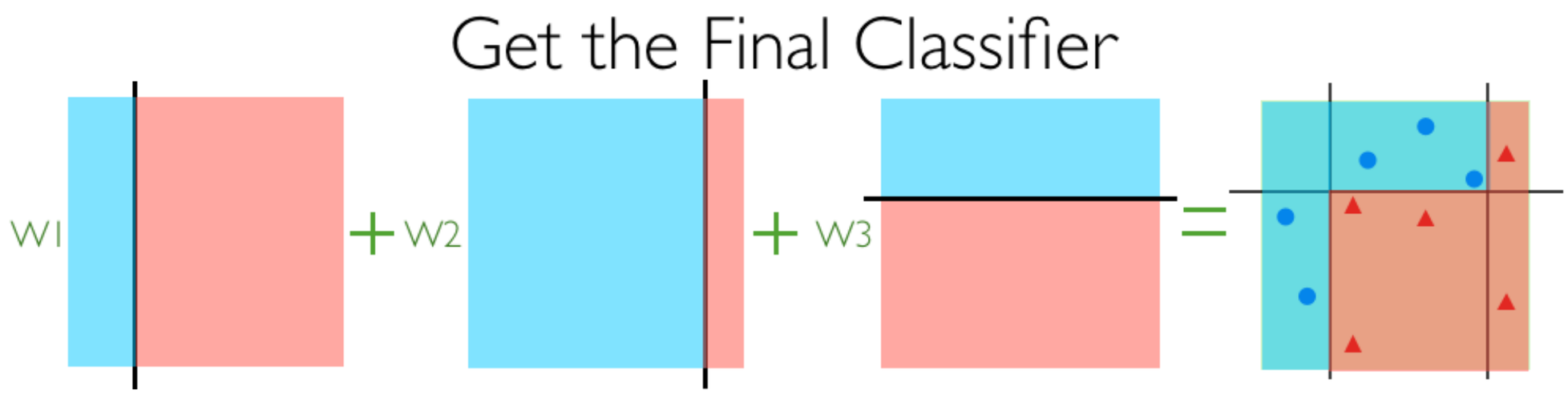
- Past the first one, each iterative boosted decision tree (classifier) is trained on the ‘same’ events. But now, the events have weights according to whether they were previously wrongly classified.



# Boosted Decision Trees

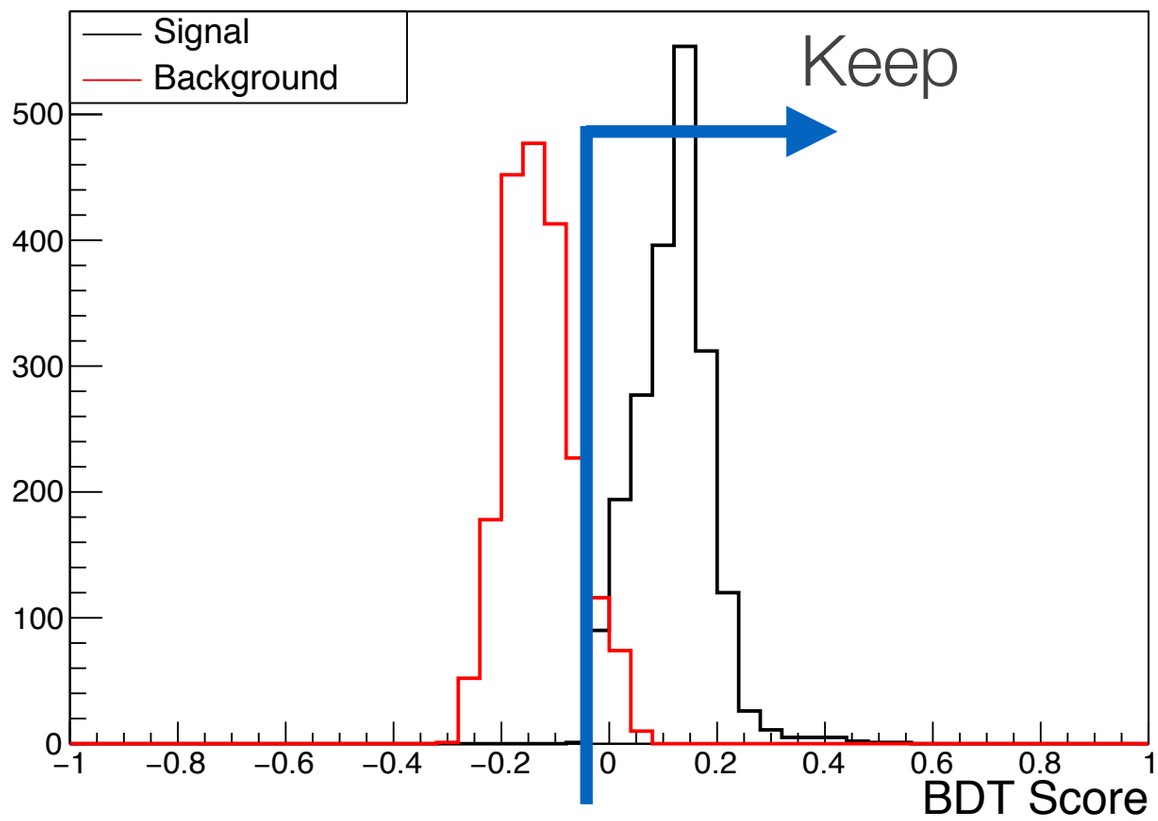
- The combined classifier is the weighted average from all trees for the different regions
- Works very well "out-of-the-box"



# Boosted Decision Tree Classifier

- After training, and hopefully testing, the BDT can generate a score when run over new data that allows signal/ background separation
  - More negative values are background
  - Place a cut at some score to get desired purity and efficiency




We are using the BDT as a classifier and want a decision about whether a **new** data event is more similar to class-A or class-B, e.g. signal or background. We use a “BDT Score” which is here the BDT decision score.

# BDT Comments

- It is common to throw an absurd number of variables into a BDT and have it signify the variables of importance. The more variables used in any supervised learning algorithm, the more difficult it is to debug when something goes wrong, e.g. user error.
- The number of nodes, variables, and depth of each tree can influence the classification outcome. But, because BDTs are generally fast to train it is often easy to tune settings by-hand
- Ensure that the variables used in training match the distribution shapes in data. Poor variable agreement will bias the BDT, and if the BDT uses many variables it can be

# Uniform Confidence Intervals

# Feldman-Cousins

# Unified Approach to Confidence Interval

- Important method for correct coverage when reporting analysis results

- It is — in my opinion — extremely useful for research when being correct is important
  - Hopefully 'being correct' is always important
  - Can be time-consuming for problems with multiple fit parameters

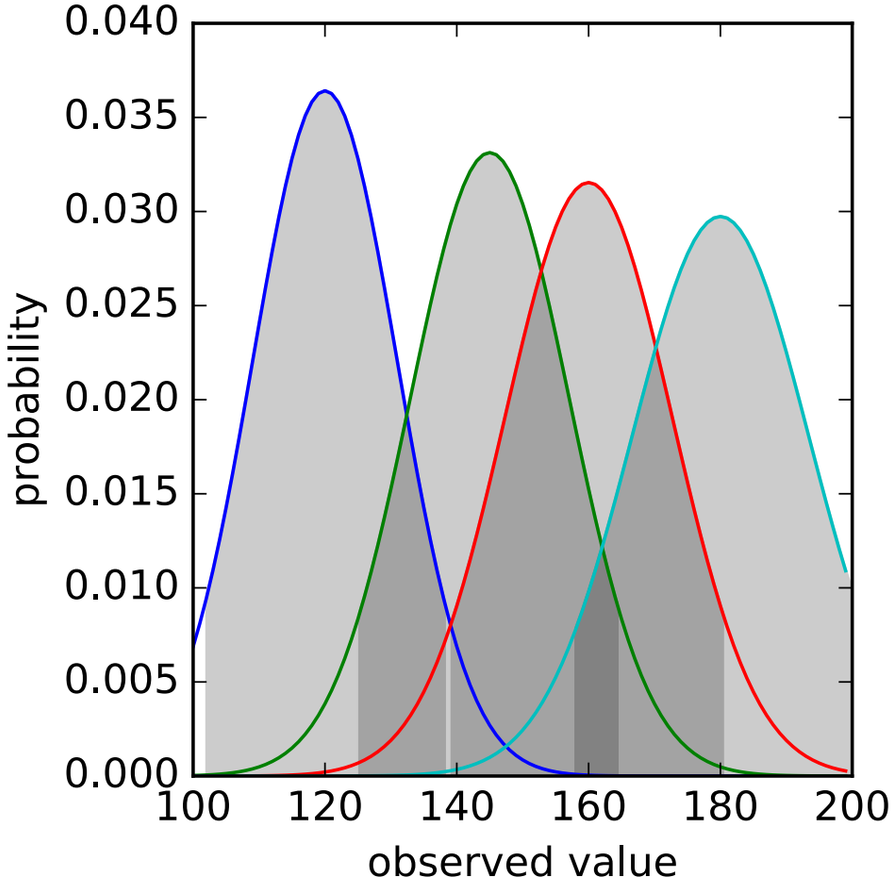
- Because simple cases are the only ones easy to do quickly, there will not be a Feldman-Cousins question on the exam

# Hypothesis rejection

- Each hypothesis will have an interval within which an observation will accept the hypothesis

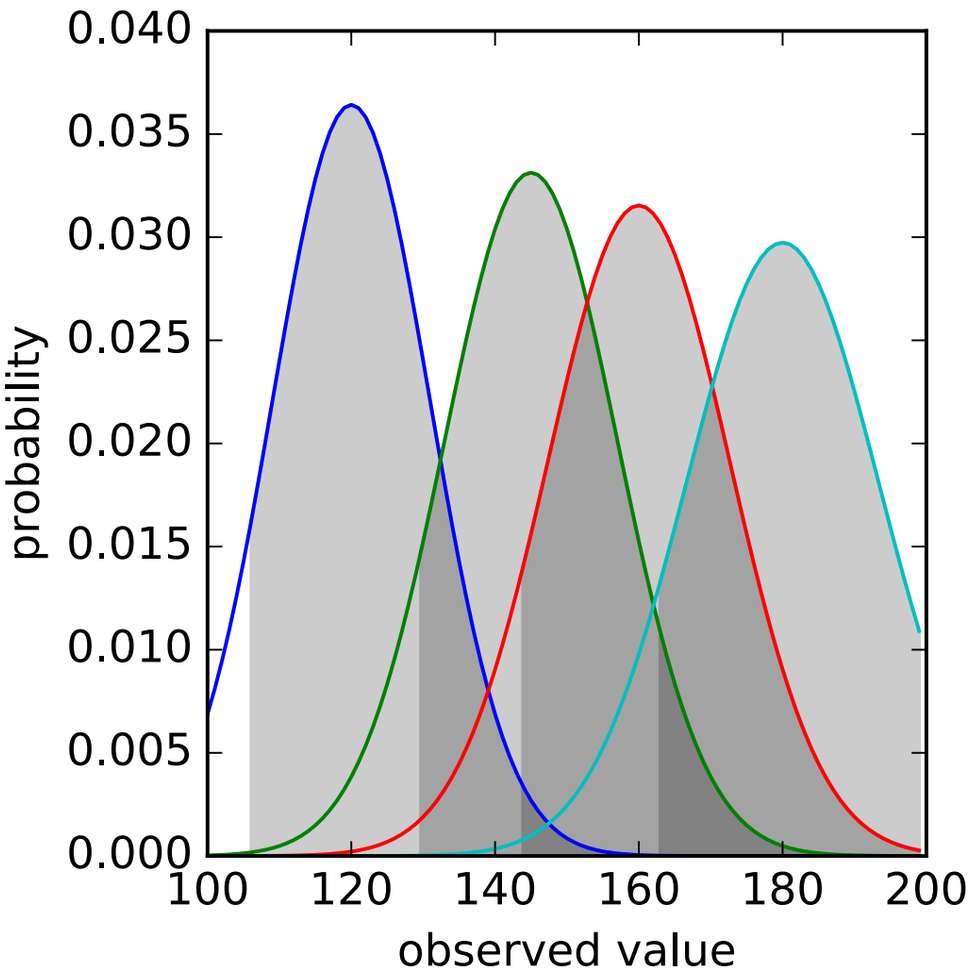
- For multiple possible true values of the parameter, the acceptance intervals can be determined

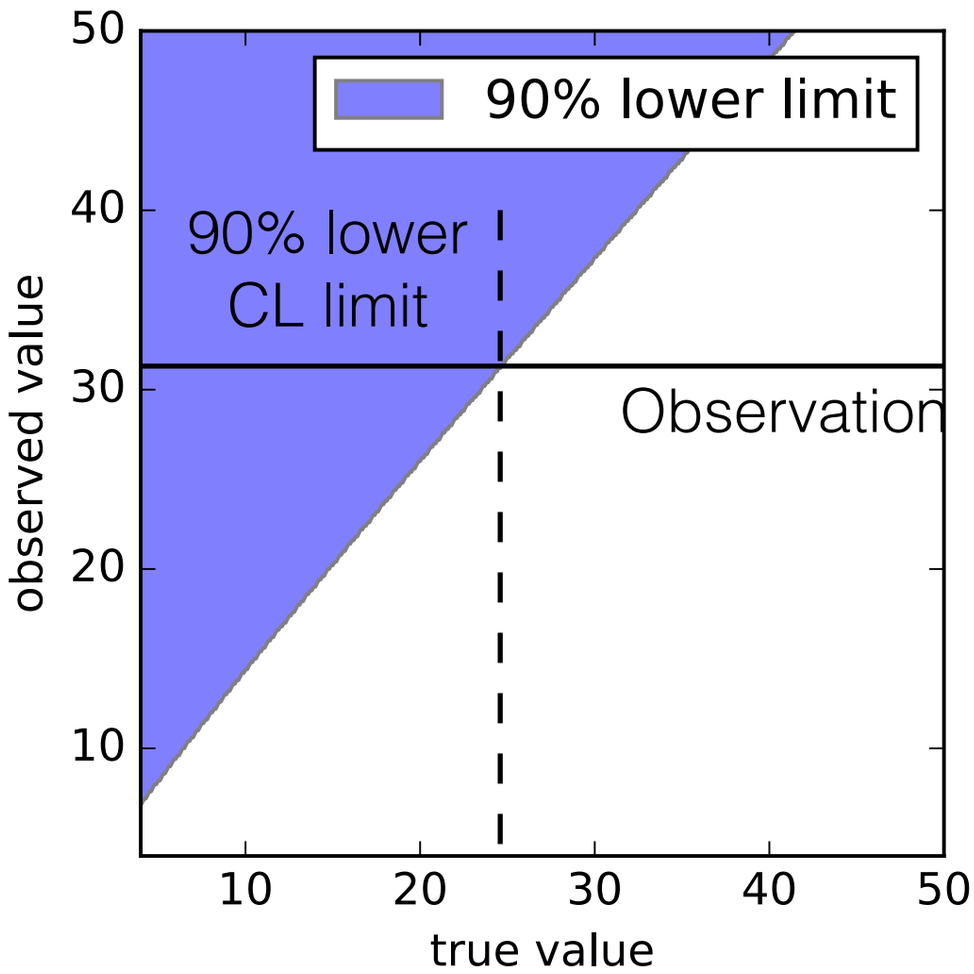
- Example on the right for 68% central interval for a few true values



# Acceptance belt

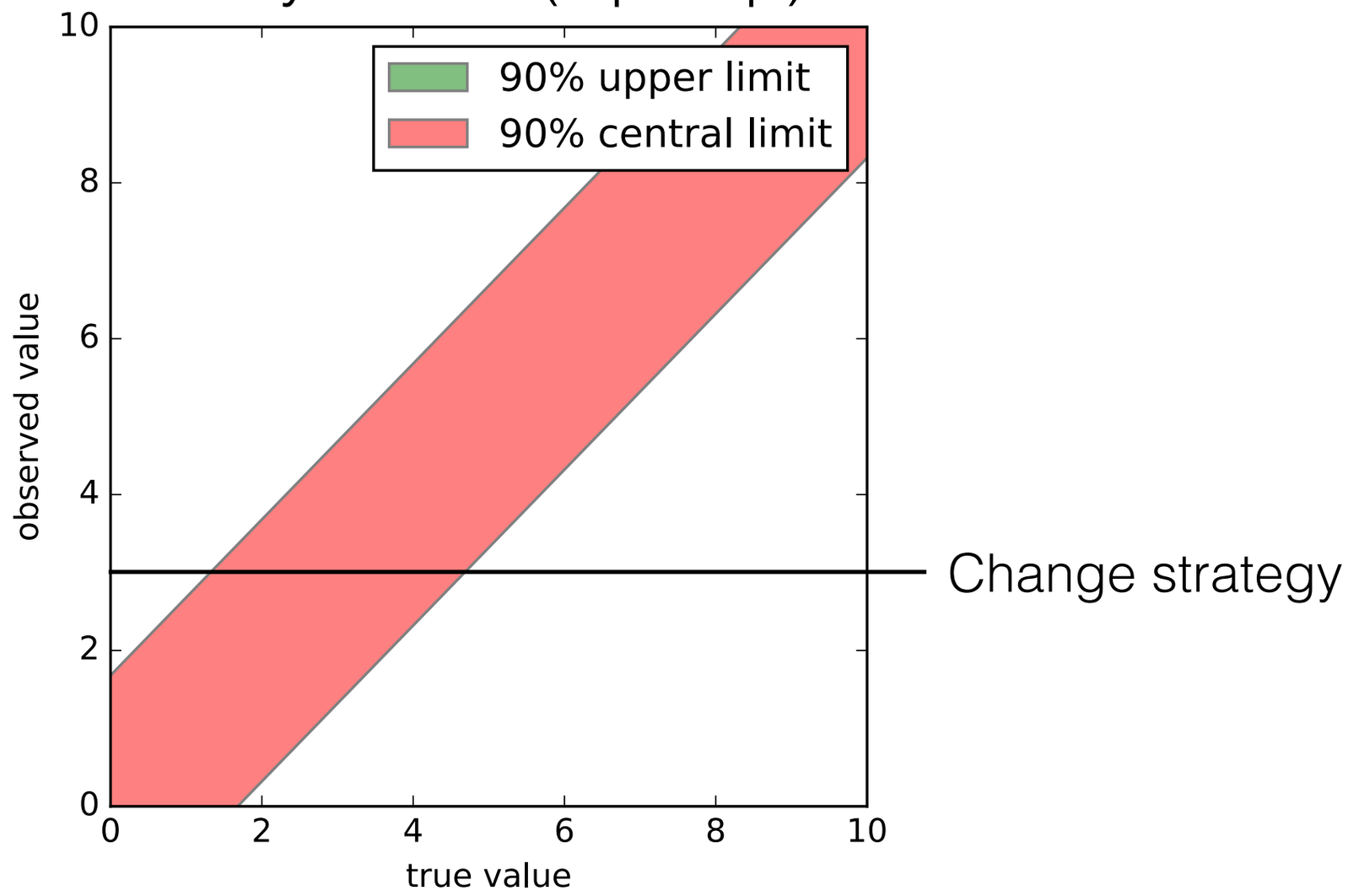
- Similarly can we produce acceptance belt for a 90% lower limit



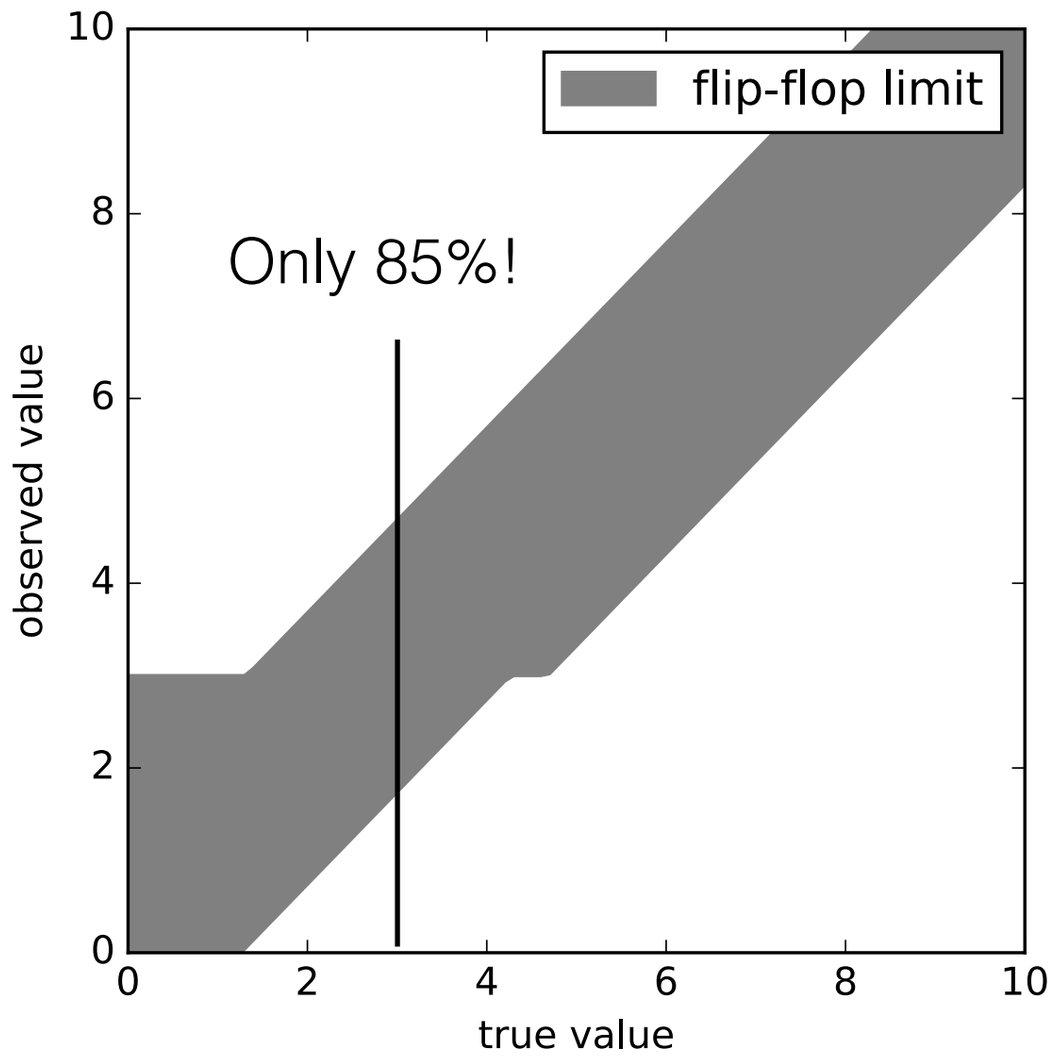
# Complication: Choosing strategy later

- Assume gaussian PDF with σ = 1, with the strategy of changing from 90% upper limits to 90% central limit if the observation is 3σ away from 0 (flip-flop)



# Complication: Choosing strategy later

- Problem: Part of the range only has 85% coverage, not the 90% that we designed the method for



# Approach

- Introduce ranking principle based on the following likelihood ratio, or rank:

$$R(\theta) = \frac{L(n|\theta)}{L(n|\theta_{\text{best}})}$$

- With the likelihood value of observing n given a true value θ, or the best fit value of the parameter $\theta_{best}$ given the dataset and any constraints on θ

- Completely rethink the construction of acceptance intervals for the acceptance belt: For a given true value θ, include values of n to the interval from highest rank R(θ) to lower, until the desired probability is reached

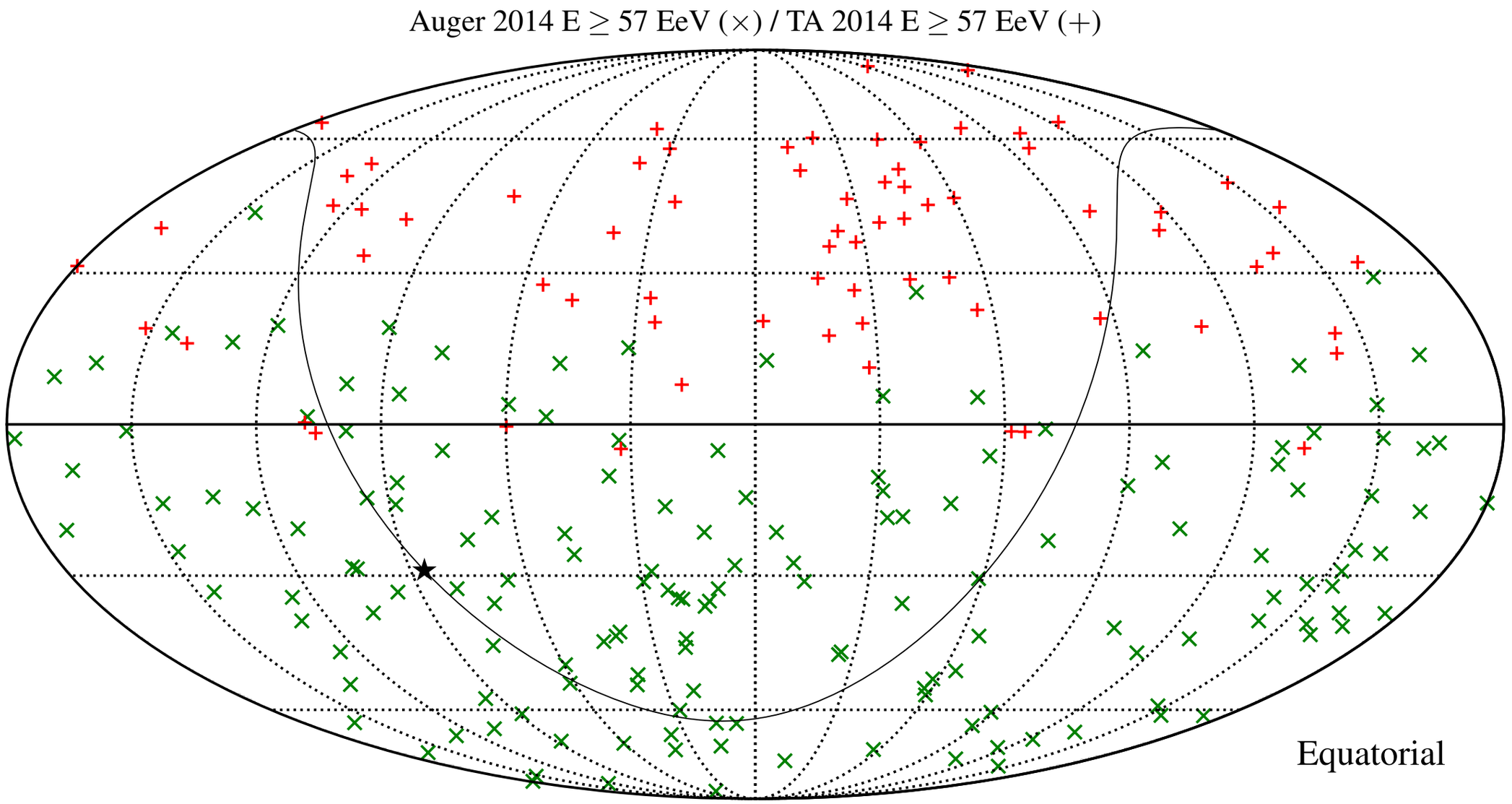
# Approach - Example

- Assume a Poisson measurement with true value θ = 1
- ‘rank’ indicates in which order the values of n are included for a 90% interval

| n | P(n\|θ=1) | $\theta_{best}$ | P(n\|$\theta_{best}$) | R | rank |
|---|---|---|---|---|---|
| 0 | 0.368 | 0 | 1 | 0.368 | 3 |
| 1 | 0.368 | 1 | 0.368 | 1 | 1 |
| 2 | 0.184 | 2 | 0.271 | 0.680 | 2 |
| 3 | 0.061 | 3 | 0.224 | 0.274 | |
| 4 | 0.015 | 4 | 0.195 | 0.079 | |
| 5 | 0.003 | 5 | 0.175 | 0.017 | |

# Auto-Correlation and Statistical Tests

# Example: Arrival Direction of Cosmic Rays




Anisotropies in the arrival directions of ultra-high energy cosmic rays (data from the observatories Telescope Array (TA) and Auger).

# Auto-Correlation

- So far, we have only looked into local excesses in individual bins.
- This method was not sensitive to the correlation between events, *e.g.* in neighbouring bins or in small clusters.
- Consider $N_{\text{tot}}$ events distributed on a sphere with position $\mathbf{n}_i$ (unit vector)
- For two events with label $i$ and $j$ ($i \neq j$) we can define an angular distance:

$$\cos\varphi_{ij} = \mathbf{n}_i \cdot \mathbf{n}_j$$

- The **cumulative two-point auto-correlation function** is defined as

$$\mathcal{C}(\{\mathbf{n}_i\}, \varphi) = \frac{2}{N_{\text{tot}}(N_{\text{tot}}-1)} \sum_{i=1}^{N_{\text{tot}}} \sum_{j=1}^{i-1} \Theta(\cos\varphi_{ij} - \cos\varphi) \qquad (2)$$

with **step function** $\Theta(x) = 1$ for $x \geq 0$ and $\Theta(x) = 0$ for $x < 0$.

➜ This expression counts the pairs of events within angular distance $\varphi$.

# Kolmogorov-Smirnov (KS) Test

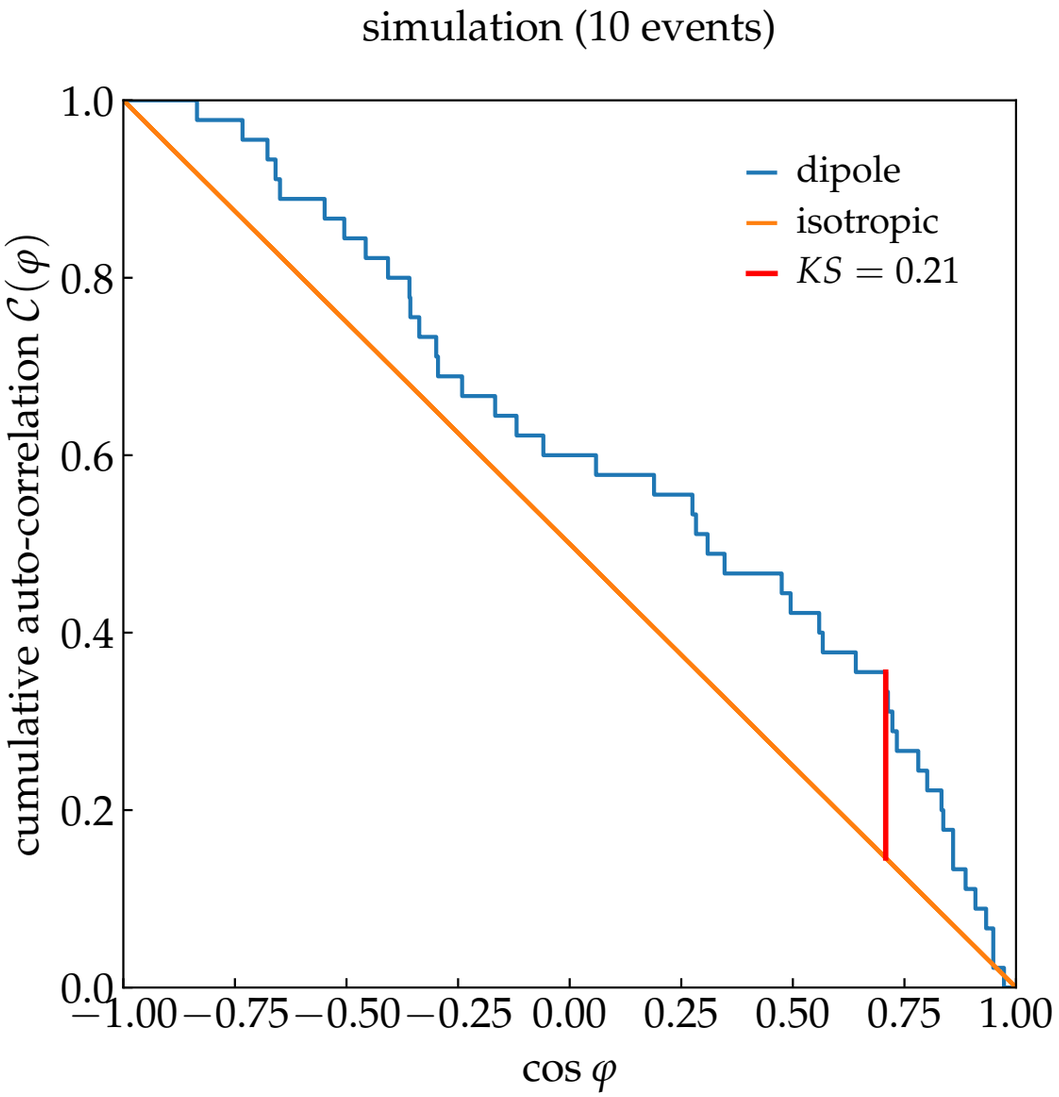
- We want to define a quantity that is a statistical measure for the difference between the empirical distribution and background distribution.
- Area between two curves?

$$\int \mathrm{d}\cos\varphi |\mathcal{C}(\{\mathbf{n}_i\},\varphi) - \mathcal{C}_{\mathrm{iso}}(\varphi)|$$

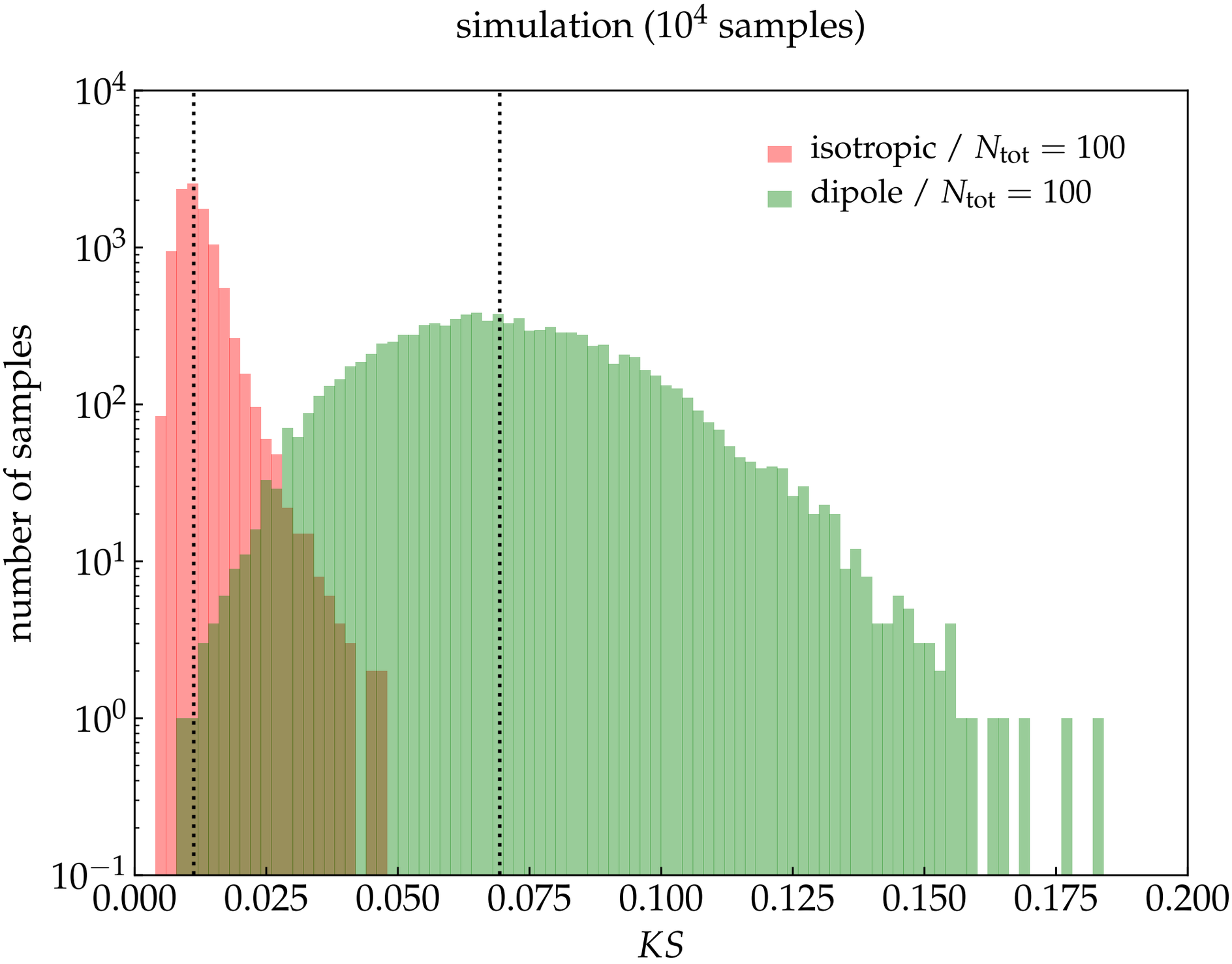
- Or, more general ($L^p$ norm)?

$$\left[\int \mathrm{d}\cos\varphi \, |\mathcal{C}(\{\mathbf{n}_i\},\varphi) - \mathcal{C}_{\mathrm{iso}}(\varphi))\,|^p\right]^{\frac{1}{p}}$$

- **Kolmogrov-Smirnov:** $p \to \infty$.



# Kolmogorov-Smirnov (KS) Test




for `python` code see : `KS_produce.py` & `KS_show.py`

# Nested Sampling

# Pure Mystic Beauty

- Nested sampling for Bayesian inference is a more recent development and can handle very complicated posterior/likelihood landscapes
- Covered just last week, so no review here…

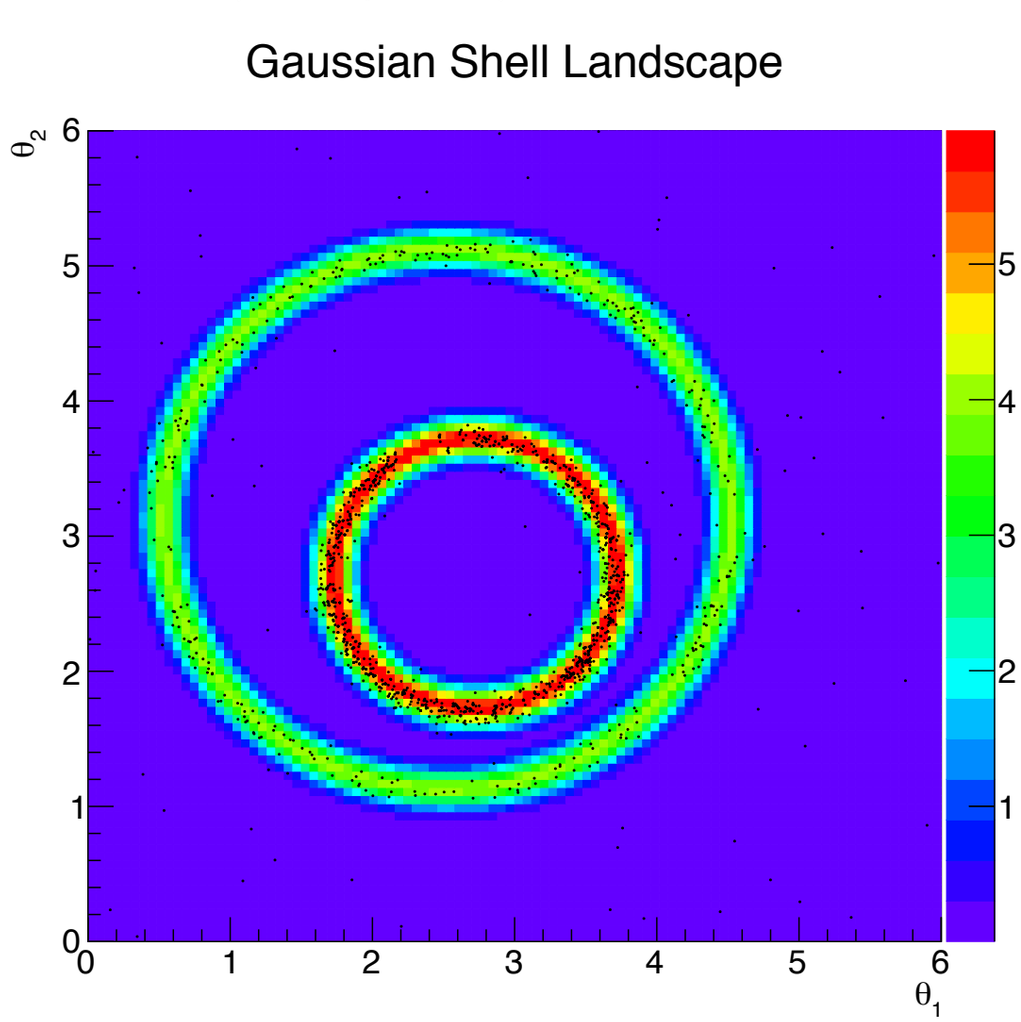
Any sufficiently advanced technology is indistinguishable from magic.
- Arthur C. Clarke

# Exercise Nested Nested Cylinder

- Using the following likelihood for the two cylinders plot the underlying likelihood and posterior distribution:

$$\mathcal{L}(\vec{\theta}) = \mathrm{circ}(\vec{\theta}; \vec{c}_1, r_1, \sigma_1) + 1.5\ \mathrm{circ}(\vec{\theta}; \vec{c}_2, r_2, \sigma_2)$$

  - $c_1$=( 2.5, 3.1) and $c_2$=( 2.7, 2.7) and $r_1$=2 and $r_2$=1



Fin

# KS-Test features

- Model being tested (and parameters) should not be drawn from the data set to which the model is being compared
- If the value of the KS statistic is out in the tails, be wary, you are dealing in low-statistics and low-likelihood regimes
  - Thankfully this suggests that the two distributions are similar
  - But, actual differences in tails of distributions are unlikely to be identified by the KS-test
- Only valid for continuous distributions
- “The distribution of the KS statistic is also not distribution-free when the dataset has two or more dimensions” -Babu & Feigelson
- Does not require binned data. Works better w/o binning

# Hypothesis Testing for KS-test

- With a specific model, commonly the null-hypothesis $H_0$ or $F_0$, we can test the max divergence with data through the EDF with the expectation from the model (or another EDF, which we'll do later)
- Math bits: The Kolmogorov-Smirnov statistic is the supremum of the point-wise EDF ($F_n(x)$) with the model CDF (F(x))

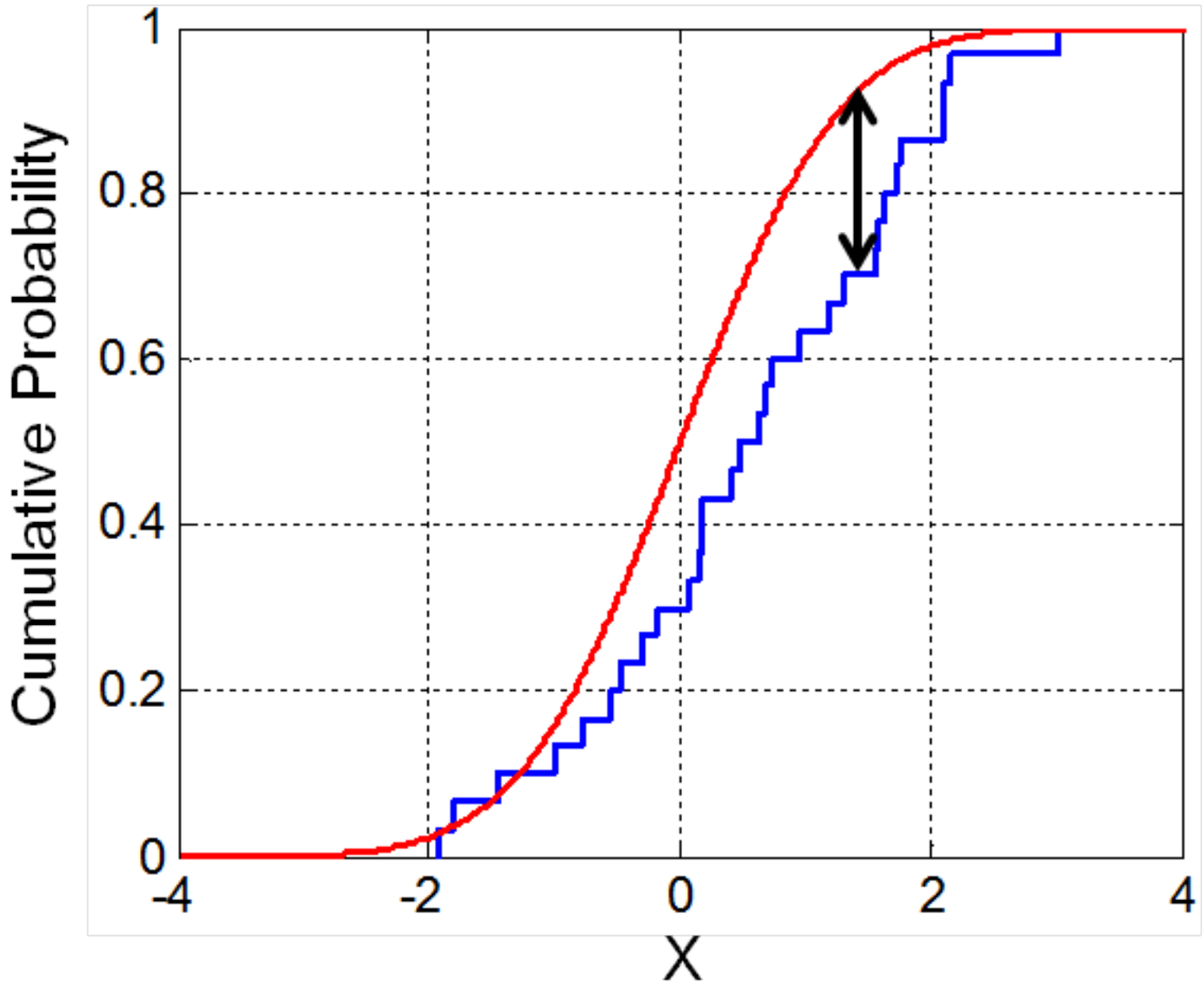
$$D_n = \sup_x |F_n(x) - F(x)|$$

- Note that the KS-test is shape-dependent. It is mostly insensitive to any normalization differences

# Graphical KS-Test

*wikipedia

- Compare the supremum, i.e. largest difference, for the two cumulative distributions




Note that both data sets can be EDFs, there is no strict requirement that both sets cannot be actual data, or sampled sets (e.g. finite statistics Monte Carlo)